Article

Large Language Model for Secure Operation of Power Systems

Yue Xiang [1,2,*], Ling Tan [3,*], Gao Qiu [1], Zhiyuan Tang [1] and Junyong Liu [1,2]

[1] College of Electrical Engineering, Sichuan University, Chengdu 610207, China; qiugscu@163.com (G.Q.); tangzhiyuan@scu.edu.cn (Z.T.); liujy@scu.edu.cn (J.L.)

[2] Clean Energy Power Systems and Equipment Key Laboratory of Sichuan Province, Chengdu 610065, China

[3] College of Electrical Engineering, Xi’an Jiaotong University, Xi’an 710049, China

* Corresponding author. E-mail: xiang@scu.edu.cn (Y.X.); tanlingee@163.com (L.T.)

Received: 16 August 2025; Accepted: 2 September 2025; Available online: 8 September 2025

ABSTRACT: The integration of large-scale renewable energy, multi-criteria operational constraints, and complex grid topologies has intensified the challenges faced by the security monitoring process within power system dispatch. Dispatch guidelines, typically expressed in natural language, are difficult for conventional algorithms to interpret and apply in real time, while general-purpose Large Language Models (LLMs) lack domain-specific knowledge, risking inaccurate or unsafe recommendations. This study proposes an LLM-based monitoring framework that integrates domain-specific prompt engineering with fuzzy evaluation to address these limitations. The framework interprets dispatch guidelines, analyzes real-time power flow data, and converts semantic assessments into quantitative safety scores, enabling closed-loop decision-making. Validation on the IEEE 14-bus system demonstrates that the optimized LLM outperforms a general LLM in accuracy, logical consistency, and stability under complex multi-standard scenarios, while reducing reliance on manual intervention. The results highlight the framework’s potential to enhance monitoring efficiency and ensure intelligent, secure power system operation.

Keywords: Large language model; Power system dispatch; Prompt engineering; Fuzzy evaluation; Safety assessment

© 2025 The authors. This is an open access article under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a sophisticated infrastructure governed by strict physical laws, the power system relies critically on its dispatching operation system for safe, stable, and reliable performance. Traditionally, power dispatching has depended on human operators who manually interpret dispatch guidelines and plans, applying professional expertise to implement operations based on the system’s real-time status. This approach is highly labor-intensive and heavily reliant on the experience of dispatchers [1]. Furthermore, the evolution of power systems has led to expanding the structure and scale of power grids. An increasing number of distributed energy resources are being integrated into the grid, leading to increasingly complex power system composition and significantly heightened operational instability. This shift poses substantial challenges to power system dispatching [2]. Moreover, the high penetration of new energy into the distribution network has introduced new challenges such as reverse power flow, insufficient regulation capacity, and deterioration of power quality [3]. To address these challenges, the government and industry have issued new dispatch guidelines, further complicating the rules. These guidelines are typically written in natural language, and traditional algorithms struggle to interpret them due to the flexibility of natural language, presenting a challenge for monitoring and dispatching systems.

Given that current dispatch systems are struggling to adapt to and solve problems related to multi-scale, multi-agent self-organization, and semantic information integration in modern power systems, recent breakthroughs in large language models (LLMs) [4] appear to offer a potential solution. They can improve the efficiency of dispatch decision-making, reduce the burden on dispatchers, and promote the intelligent management of modern power systems. Models such as Transformer, Llama [5], and ChatGPT [6] have mastered the deep structure and context of language through extensive pre-training, enabling them to understand and follow complex instructions. When finely tuned, LLMs can perform specific tasks at or above human-level efficiency, especially with prompt engineering [7], enhancing their adaptability to new tasks without requiring extensive retraining. This flexibility greatly benefits complex decision-

https://doi.org/10.70322/sesr.2025.10005

making and human-machine collaboration. LLMs' powerful natural language processing capabilities are critical for efficient collaboration between humans and machines. Their generalization ability minimizes the need for new models tailored to different scenarios, demonstrating their versatility and wide applicability [8]. These characteristics make LLMs particularly well-suited for tackling the semantic interpretation and decision-making challenges inherent in modern power system dispatch.

While emerging, LLMs applications in power systems have predominantly focused on specific sub-tasks. Examples include wind and load forecasting [9,10], fault diagnosis [11], dispatching [12], model generation from descriptions [13], algorithm synthesis [14], and regulatory information retrieval [15]. While valuable, these applications remain largely offline or auxiliary in nature. In contrast, the application of LLMs to core real-time decision-making for secure operation, which requires actions to respect strict physical, operational, and reliability constraints, remains largely unexplored. This research is among the first to embed LLMs directly into the real-time dispatch loop. This framework translates live system data and operational rules into quantitative safety scores and actionable guidance, creating a decision-support feedback loop that informs secure dispatch in real time, thus enhancing the security and reliability of system operation.

However, applying LLMs to this core function presents significant challenges. General-purpose LLMs, such as GPT-4 [16], are trained on broad internet corpora and lack detailed knowledge of power system physics, operating procedures, and security constraints. Without careful adaptation, they may generate misleading or unsafe recommendations through oversimplified reasoning or erroneous interpretations of grid states. Moreover, real-time operation imposes strict requirements on latency, reliability, and interpretability that general-purpose LLMs are not designed to meet. Addressing these barriers requires structured prompt design, domain adaptation, and mechanisms for verifying model outputs against operational criteria.

In summary, this paper is conducted under the assumptions that the power system operates without abnormal disturbances in power flow data and that safety thresholds remain static. Based on these assumptions, this study is motivated by the following research questions:

- A gap exists in leveraging LLMs for power system dispatch, as general-purpose models lack specialized domain knowledge to accurately interpret dispatch guidelines and grid operation rules.
- There is a need to improve the reliability of real-time security assessment under multi-semantic forms of safety standards, ensuring consistent and trustworthy decision-making.
- Existing dispatching approaches remain inefficient and heavily dependent on manual expertise, highlighting the necessity of an intelligent framework that enhances efficiency and reduces operator reliance.

In response, this paper harnesses the potential of LLMs and introduces a novel LLMs-based framework for the secure operation of power systems. The contributions are summarized as follows:

- A monitoring framework is proposed that integrates domain-specific prompt engineering with fuzzy evaluation, enabling the transformation of natural-language dispatch guidelines into quantitative security assessments.
- A real-time decision-making mechanism is developed that supports multi-criteria security evaluation and closed-loop monitoring, where the system not only detects security violations but also proposes actionable dispatch responses, thereby improving the accuracy, logical consistency, and stability of system operation under complex conditions.
- The framework enhances dispatching efficiency and reduces operator dependence, providing a practical and forward-looking pathway toward an autonomous and intelligent "dispatch brain" for power systems.

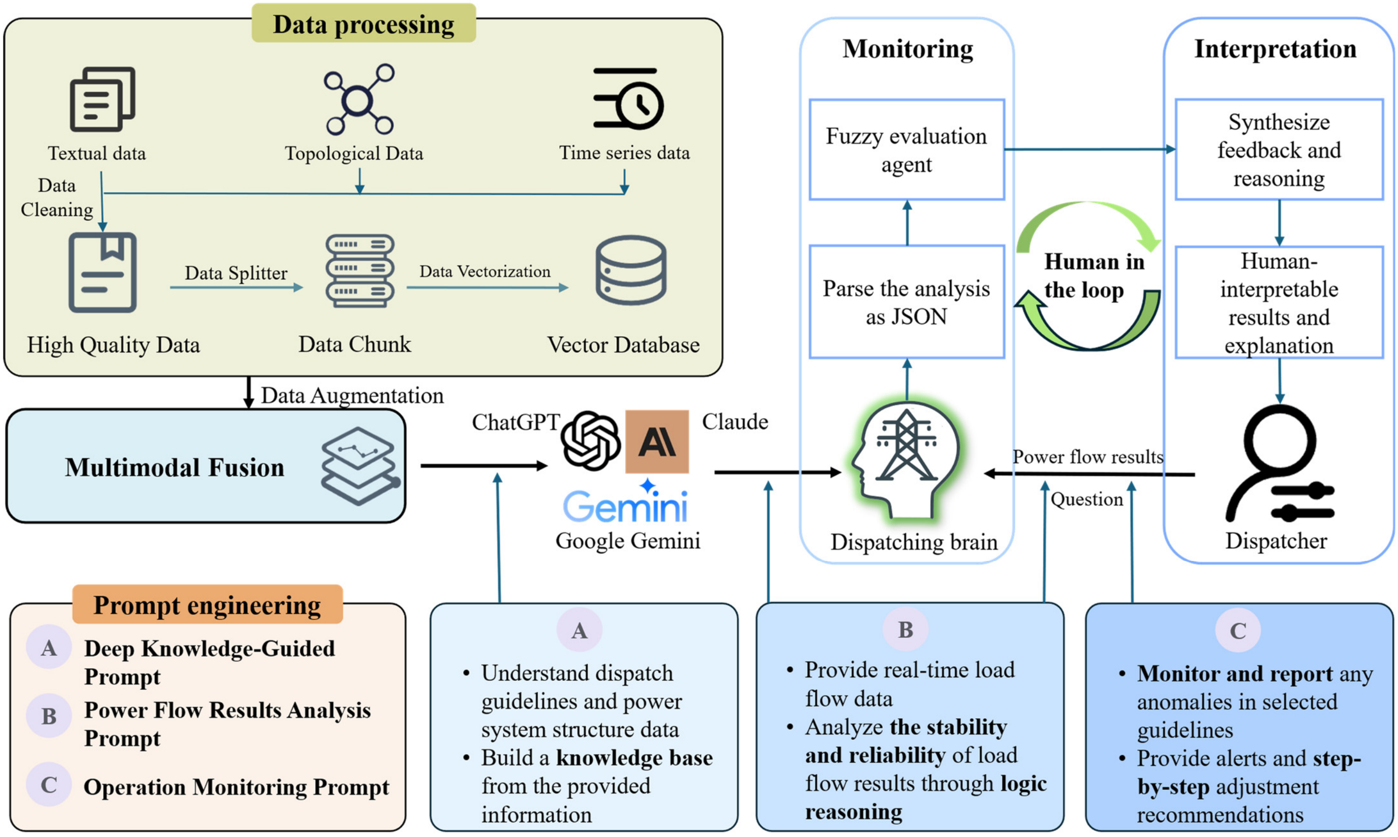
The rest of the paper is organized as follows. Section 2 presents the proposed LLM-based monitoring framework, including its overall architecture, prompt design, and fuzzy evaluation strategies. Section 3 reports the simulation results on the IEEE 14-bus system and evaluates the framework against general-purpose models. Section 4 discusses the findings and concludes the study.

2. Materials and Methods

The integration of advanced technologies, such as LLMs, has introduced new possibilities for enhancing the efficiency and safety of power system dispatch. By leveraging natural language processing capabilities in conjunction with data-driven analysis, these models offer real-time decision support and risk assessment for dispatchers. This not only alleviates the operational burden on human operators but also improves the responsiveness and adaptability of the dispatch process. The following section will provide a detailed exposition of the proposed methodological framework, emphasizing the key components of the LLM-based approach and how the integration of semantic understanding with fuzzy evaluation contributes to the secure operation of the power system.

2.1. Overall Framework

This study proposes a large language model-based intelligent early warning method for power systems. This method leverages the powerful natural language understanding and generation capabilities of large semantic models, combined with fuzzy evaluation, to convert the model's response into quantifiable data, thereby improving the real-time accuracy and safety of power system warnings. The overall framework consists of the following key modules: data processing module, power flow calculation module, semantic model module, and fuzzy evaluation processing module. These modules collaborate through carefully designed interfaces to ensure smooth data flow and efficient processing within the system, as shown in Figure 1.



Data processing
Textual data
Topological Data
Time series data
Data Cleaning
Data Splitter
Data Vectorization
High Quality Data
Data Chunk
Vector Database
Data Augmentation
Multimodal Fusion
ChatGPT
Claude
Gemini
Google Gemini
Monitoring
Fuzzy evaluation agent
Parse the analysis as JSON
Dispatching brain
Human in the loop
Interpretation
Synthesize feedback and reasoning
Human-interpretable results and explanation
Dispatcher
Power flow results
Question
Prompt engineering
A Deep Knowledge-Guided Prompt
B Power Flow Results Analysis Prompt
C Operation Monitoring Prompt
A
• Understand dispatch guidelines and power system structure data
• Build a knowledge base from the provided information
B
• Provide real-time load flow data
• Analyze the stability and reliability of load flow results through logic reasoning
C
• Monitor and report any anomalies in selected guidelines
• Provide alerts and step-by-step adjustment recommendations

Figure 1. Framework overview.

2.1.1. The Data Processing Module

The Data Processing Module interface first receives raw data from multiple sources, including dispatch guidelines [17], professional literature [18], and standardized grid structure documents [19,20], which are typically in the form of natural language text. Therefore, it is necessary to perform preprocessing and vectorization on these data. With the support of the interface, the data processing module transforms this textual data into a structured vector format. Specifically, this process includes the following steps:

Step 1: Data Cleaning: Remove unnecessary characters from the text and standardize the character format to cleanse and standardize the data format.

Step 2: Text Normalization: Simplify and standardize the text data by removing stop words, applying stemming and lemmatization techniques, and merging synonyms.

Step 3: Tokenization: Use tokenization techniques to process the text, splitting continuous text strings into independent, recognizable units.

Step 4: Text Data Augmentation: Enhance the data's expressiveness through word vector-based similarity techniques.

Step 5: Data Vectorization: Use word vector libraries to obtain word vectors and calculate the average of these vectors to derive the vector representation of the entire text.

The above processing steps ensure the standardization and consistency of the output data, laying a solid foundation for subsequent semantic model parsing.

2.1.2. Power Flow Calculation Module

The Power Flow Calculation Module calculates the power flow results of the power system according to Equation (1)

$$Y = [V_i^t, P_{ij}^t, Q_{ij}^t] \tag{1}$$

where V_i^t represents the voltage at bus i at time t, P_{ij}^t represents the active power of the branch from bus i to bus j at time t, and Q_{ij}^t represents the reactive power of the branch from bus i to bus j at time t. Since the power flow results are presented in a table-like format, with the content mainly consisting of related symbols, letters, and numbers, they are saved as a txt file and passed to the semantic model in LaTeX format for learning, which improves the model's accuracy in parsing numbers and letters [21].

2.1.3. Monitoring and Safety Assessment Module Based on LLM

The design of the LLM aims to effectively leverage the natural language processing capabilities of the LLM for semantic understanding and knowledge internalization of the processed dispatch guidelines. It analyses system power flow data based on the knowledge base and provides early warnings. Through this interface, the LLM receives the vectorized dispatch guidelines set $S = \{s_1, s_2, \ldots, s_n \mid n \in N\}$ and dispatch criteria set $X = \{x_1, x_2, \ldots, x_m \mid m \in M\}$, generating the corresponding knowledge base G. During the dispatch process, the LLM analyses the system's power flow results in real time and provides early warnings and decision support based on the dispatch criteria selected by the dispatcher. The analysis results are output through the interface in JSON format, and a comprehensive system safety score is provided based on fuzzy evaluation for the dispatcher's reference.

2.2. Prompt Designing

In this study, the key to designing early warning prompts for power systems lies in building an information architecture that guides the LLM to learn dispatch guidelines and the grid structure of power systems, enabling it to comprehensively assess system status and generate warnings when receiving real-time power flow data, as shown in Figure 2. Unlike traditional data-driven methods, this approach helps the model establish a systematic understanding by inputting dispatch guidelines and typical system parameter information, enhancing its accuracy and adaptability in real-world early warning applications.

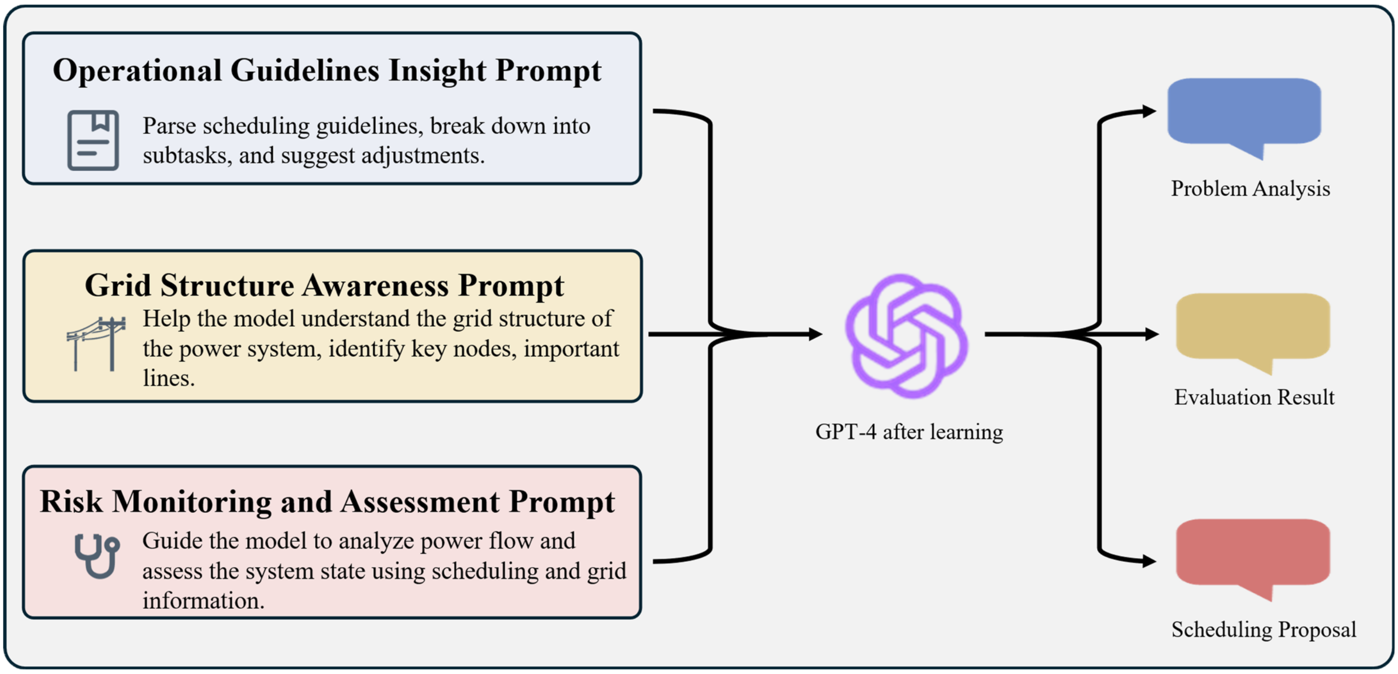

Operational Guidelines Insight Prompt
Parse scheduling guidelines, break down into subtasks, and suggest adjustments.
Grid Structure Awareness Prompt
Help the model understand the grid structure of the power system, identify key nodes, important lines.
Risk Monitoring and Assessment Prompt
Guide the model to analyze power flow and assess the system state using scheduling and grid information.
GPT-4 after learning
Problem Analysis
Evaluation Result
Scheduling Proposal

Figure 2. Prompt Designing.

First, the prompt design of the model is based on the interpretation of power dispatch guidelines. These guidelines are critical standards for the safe operation of power systems, so we designed a prompt framework that guides the model to learn and apply them gradually. This framework breaks down complex dispatch procedures into multiple sub-tasks and constraints, allowing the model to apply the guidelines in different operational scenarios flexibly. For example, the model uses prompt instructions to identify whether the load on a transmission line is approaching its safety threshold and generates corresponding adjustment recommendations based on the guidelines. This prompt method ensures that the model accurately grasps the operational requirements of the dispatch guidelines and can respond quickly in emergencies [22].

In addition, to enhance the model's understanding of the power grid structure, we designed a prompt framework based on grid structure learning. The complexity of grid structures is crucial to system stability, and the model must be able to identify critical nodes, important lines, and potential bottleneck areas within the system. By providing prompt information that describes the system's topology, the model comprehensively understands the electrical connections between nodes and their priorities in system operation. For instance, the model is prompted to analyze whether voltage fluctuations at key nodes could trigger system instability when power flows abnormally. The prompt design also requires the model to assess potential chain reactions caused by fragile links during peak load periods or in sudden situations. This design helps the model effectively identify potential risks in the grid and generate timely early warning recommendations.

In practical applications, real-time power flow data is one of the most dynamic factors in power systems. To address this, we designed specific real-time data analysis prompts that assist the model in integrating real-time power flow data with the dispatch guidelines and grid structure information it has learned for comprehensive analysis and evaluation. After receiving real-time power flow data, the model is prompted to analyze the load and voltage status of key lines and nodes and assess whether they exceed safety thresholds based on dispatch guidelines. For example, when the model detects that the load on a line is approaching its maximum capacity, it generates corresponding warnings and adjustment strategies based on prompts to avoid potential system failures. This prompt design ensures the model has real-time responsiveness in complex power system environments, allowing it to generate effective risk warnings based on dynamic data.

The prompt design in this study not only helps the model extract key information from power dispatch guidelines and grid structure data but also ensures that the model can quickly analyze system status and generate warnings when receiving real-time power flow data. This prompt design approach significantly enhances the model's adaptability, accuracy, and reliability in real-world power system applications, providing strong technical support for power system early warning.

2.3. Multi-Criteria Selection and Early Warning Strategies Based on Fuzzy Evaluation

In LLM-based power system dispatch, multi-criteria analysis and fuzzy evaluation are key to managing system complexity and uncertainty. Multi-criteria analysis helps dispatchers weigh the importance of various indicators, such as line overload, voltage violations, frequency stability, and power factors, leading to more informed decisions. These criteria, developed by experts, reflect critical aspects of power system stability and safety.

When conducting multi-criteria analysis, dispatchers first need to select relevant criteria based on the current state and operating conditions of the system. For example, when the system is under high load conditions, line overload may be the most critical concern, while under low load conditions, voltage violations might be more crucial. By dynamically selecting criteria, dispatchers can focus more specifically on the key risks of the system. Once the relevant criteria are determined, the LLM will sequentially analyze the power flow results based on the selected criteria and assess whether the system is safe under each criterion.

When analyzing power flow results, the LLM evaluates the system's safety by assessing key parameters like line load and voltage levels, using its built-in rule base. For example, if it detects "overload" or "voltage anomaly", it generates a safety assessment. The LLM also provides dispatch suggestions based on the current system state and established rules, such as recommending load shedding or redistributing generator outputs. These recommendations are presented in natural language, offering guidance to dispatchers for initial decision-making.

Following the model's safety evaluations, a fuzzy evaluation is conducted to quantify the overall system safety. These numerical results provide dispatchers with more precise references, helping them make more effective decisions. Since the LLM's responses are given in natural language, fuzzy evaluation translates these complex semantic results into numerical forms, enabling reliable system safety assessments. Certain states may be described ambiguously in power systems, such as “slightly overloaded” or “severely overloaded”. These descriptions are converted into numerical values through fuzzy logic, offering a more objective reflection of the system’s actual state. Considering the variability in the language format of the model’s responses, semantic parsing is necessary to categorize similar semantic judgments into the same class. A method based on sentence embeddings and semantic similarity calculations is primarily employed here. First, the Sentence-BERT model is used to generate vector representations of both the sentences produced by the LLM and predefined evaluation template sentences according to Equation (2). These vector representations then calculate the cosine similarity between sentences, matching the LLM’s natural language outputs with predefined scoring templates.

$$v(S_i) = SBERT(S_i) \tag{2}$$

where S_i represents the output judgment statement of the LLM based on the dispatch criteria’s evaluation of the power flow results, and $v(S_i)$ represents the vector representation of the sentence S_i generated by the Sentence-BERT model.

To accurately reflect the severity of incidents, this study predefines a set of template sentences and their corresponding safety scores for each dispatch criterion as shown in Equation (3) [23]. The design of these templates is informed by regulations on power safety emergency response and investigation [24], as well as by operational guidelines. These templates describe the various safety states of the power system.

$$r(T_k) = \begin{cases} 0.2 & T_k = \text{"}\textit{A particularly serious accident occurred under this guideline.}\text{"} \\ 0.4 & T_k = \text{"}\textit{A major accident occurred under this guideline.}\text{"} \\ 0.6 & T_k = \text{"}\textit{A relatively major accident occurred under this guideline.}\text{"} \\ 0.8 & T_k = \text{"}\textit{A general accident occurred under this guideline.}\text{"} \\ 1.0 & T_k = \text{"}\textit{No accidents occurred under this guideline.}\text{"} \end{cases} \tag{3}$$

Based on the output of the semantic model, we can calculate the semantic similarity between the sentence generated by the semantic model and each template sentence, which is the cosine similarity between the two sentence vectors.

$$sim(S_i, T_k) = \frac{v(S_i) \cdot v(T_k)}{\|v(S)\| \|v(T_k)\|}, \forall k \in \{1, \ldots, K\} \tag{4}$$

where $\|v(S_i)\|$ and $\|v(T_k)\|$ are the Euclidean norms of vectors $v(S_i)$ and $v(T_k)$, respectively, used to normalize the dot product result to ensure the similarity value falls between −1 and 1.

$$T_{k*} = \arg\max_{T_k}(sim(S_i, T_k)) \tag{5}$$

$$sim_i^{\max} = \max_{T_k} \sin(S_i, T_k) \tag{6}$$

where T_{k^*} represents the predefined sentence T_k that maximizes the value of $sim(S_i, T_k)$; $sim_i^{\max}$ represents the maximum semantic similarity between S_i and all predefined sentences T_k.

It should be noted that the reliance on sentence-level similarity scoring based on Sentence-BERT embeddings with cosine similarity may face challenges in low-resource or semantically ambiguous cases where subtle linguistic variations are not fully captured by embeddings [25,26]. However, in our framework, this potential limitation does not compromise the correctness of the overall monitoring process, since the fuzzy evaluation mechanism aggregates similarity scores into a multi-level membership function, which reduces the impact of occasional misalignment.

To quantify the overall safety state of the system, we combine the membership degree of the template sentences with the semantic similarity between sentences. We then calculate the comprehensive safety score under the selected guidelines M from the set of scheduling guidelines and normalize it:

$$p(S) = \frac{\sum_{i=1}^{a} sim_i^{\max} \cdot r(T_{k*})}{\sum_{i=1}^{a} sim_i^{\max}} \tag{7}$$

where, $p(S)$ represents the comprehensive safety score, reflecting the overall safety of the system in its current state. a represents the number of selected scheduling guidelines. Here, a fuzzy set is defined as follows:

- Normal state: The system operates well, and all safety indicators are within normal ranges.
- Warning state: There are minor issues in the system, but they do not pose an immediate significant impact on safety.
- Emergency state: The system has encountered serious problems that require quick action, though the system is not in immediate danger of collapse.
- Critical state: The system is in a highly risky state, which could lead to system collapse and requires immediate intervention.

The membership function is defined as follows:

$$\mu_{normal}(p(S)) = \begin{cases} 1 & if\ p(S) \geq 0.85 \\ \dfrac{p(S) - 0.7}{0.15} & if\ 0.7 \leq p(S) < 0.85 \\ 0 & if\ p(S) < 0.7 \end{cases} \tag{8}$$

$$\mu_{warning}(p(S)) = \begin{cases} 0 & if\ p(S) \leq 0.5 \\ \dfrac{p(S) - 0.5}{0.2} & if\ 0.5 < p(S) \leq 0.7 \\ \dfrac{0.8 - p(S)}{0.1} & if\ 0.7 < p(S) \leq 0.8 \\ 0 & if\ p(S) > 0.8 \end{cases} \tag{9}$$

$$\mu_{emergency}(p(S)) = \begin{cases} 0 & if\ p(S) \leq 0.3 \\ \dfrac{p(S) - 0.3}{0.1} & if\ 0.3 < p(S) \leq 0.4 \\ \dfrac{0.6 - p(S)}{0.2} & if\ 0.4 < p(S) \leq 0.6 \\ 0 & if\ p(S) > 0.6 \end{cases} \tag{10}$$

$$\mu_{critical}(p(S)) = \begin{cases} 1 & if\ p(S) \leq 0.2 \\ \dfrac{0.4 - p(S)}{0.2} & if\ 0.2 < p(S) \leq 0.4 \\ 0 & if\ p(S) > 0.4 \end{cases} \tag{11}$$

Here, the membership functions for the fuzzy evaluation are designed as piecewise linear functions, as given in Equations (8)–(11) and Table 1. This design choice differs from parametric fuzzy models such as the Takagi–Sugeno type [27], since the objective of the framework is not nonlinear approximation but the provision of transparent and operationally interpretable safety scores. Piecewise linear functions establish explicit threshold boundaries and gradual transitions between safety states, consistent with practical engineering standards, including voltage limits, line overload thresholds, and incident severity levels specified in dispatching guidelines.

Moreover, this formulation enables efficient translation of linguistic judgments into quantitative scores. For example, states described as "slightly overloaded" or "severely overloaded" can be directly mapped to numerical values through linear membership scaling, which avoids the complexity of parametric fuzzy models and ensures computational tractability in real-time applications. Such properties are particularly important in security monitoring, where interpretability and responsiveness are more critical than the ability to approximate nonlinearities [28,29].

Table 1. Power System State Safety Score Levels.

	Normal	**Warning**	**Emergency**	**Critical**
$\mu(p(S))$	[0.8,1.0]	[0.6,0.8)	[0.4,0.6)	[0,0.4)

Using these membership functions, we can determine the system's degree of membership in each fuzzy set and conduct a comprehensive safety evaluation based on the principle of maximum membership. The thresholds for each safety level are derived from historical data and expert knowledge, but can be adjusted in practical applications to meet specific safety and operational requirements.

3. Results

In this study, GPT-4 was selected as the baseline large language model for experimental validation. Two configurations were considered: a general GPT-4 model without domain-specific adaptation and an optimized GPT-4 model enhanced through prompt engineering and the incorporation of power system documents. This design ensures that any observed improvements can be attributed to integrating domain knowledge and structured prompts rather than differences in model architecture.

To accurately quantify the effectiveness of the proposed approach, we employed a unified set of evaluation metrics to compare the performance of the general model and its optimized variant, focusing on authenticity, logical consistency, stability, and expressive capability.

3.1. Evaluation Metrics System

To validate the feasibility of this method and consider the specificity of the results generated by the semantic model, we adopted four metrics for analysis: authenticity, logical consistency, stability, and expressiveness [14,30]. Authenticity refers to the accuracy and reliability of the content generated by the model. In power system early warning, the model's output must be based on real data; otherwise, it may affect the decision-making process and operational efficiency, potentially leading to severe system failures [31]. To quantify authenticity, we compare the model's output with known standard answers, using the percentage of correct responses out of the total responses to measure accuracy, as shown in Equation (12).

$$F\_a = \frac{NCR}{TNR} \times 100\% \tag{12}$$

where, NCR represents the number of correct responses, and TNR denotes the total number of responses.

Logical consistency measures the coherence of the output in terms of semantics and reasoning, ensuring that the content generated by the model is logically accurate and based on correct, real-time, and non-misleading information [32,33]. This metric is difficult to quantify using a mathematical formula, so we chose GPT-4 to score the logical consistency of the output text. The scoring focuses on three aspects: the structure and order of information, sentence relevance, and keyword coherence, with a maximum score of 1 point.

Stability refers to the model's ability to generate consistent results when faced with similar or identical inputs [34,35]. In this method, the model must ensure consistency in the output results across multiple instances of the same or similar inputs, which can be measured by the cosine similarity between the output results. Additionally, divergence is introduced to assess variations in the output distribution of the model. If the model produces significantly different results for similar inputs, it indicates poor stability; conversely, if the changes are minimal, it suggests strong robustness.

$$S\_t = \frac{1}{k-1}\sum_{i=1}^{k-1}\left(\frac{o_i \cdot o_{i+1}}{\|o_i\|\|o_{i+1}\|}\right) + D_{KL} \tag{13}$$

where k represents the number of outputs generated under similar inputs; o_i represents the semantic vectors of the i-th. D_{KL} is used to measure the difference between two output distributions. The greater the difference between the output distributions, the larger the divergence value KL, indicating poorer output stability.

Expressiveness refers to the model's ability to convey information clearly and accurately. In this method, the model not only needs to provide correct reasoning but also must deliver the information to the operator in a concise and understandable manner. This can be quantified through readability score, information redundancy, and key information density, as shown in Equations (14)–(17) [36].

$$S_{SMOG\_ori} = 1.0430 \times \sqrt{P \times \frac{30}{S}} + 3.1291 \tag{14}$$

$$S_{SMOG} = \frac{(S_{SMOG\_ori} - \min(S_{SMOG\_ori}))}{\max(S_{SMOG\_ori}) - \min(S_{SMOG\_ori})} \tag{15}$$

$$R\_e = 1 - \frac{u\_f}{S} \tag{16}$$

$$D_{key} = \frac{I_{key}}{I_{total}} \times 100\% \tag{17}$$

where S_{SMOG_ori} represents the readability score, where a higher score indicates that the text is more complex and less readable. S_{SMOG} denotes the normalized readability score, constrained between 0 and 1. P refers to the number of polysyllabic words in the generated text. S represents the number of sentences in the generated text. R_e represents information redundancy. u_f denotes the number of different sentences in the text. D_{key} represents key information density. I_{key} represents the number of key information words.

3.2. Scenario Analysis

In this section, we selected the IEEE 14-bus system for analysis to verify the feasibility of this method. Bus voltage limits adhered to those in the benchmark dataset, while the rated capacities of all transmission lines were uniformly set to 100 MVA to eliminate ambiguity in thermal ratings. To enable a controlled, stepwise evaluation aligned with operational practice, three scenarios were adopted that reflect an escalating monitoring scope and increasing interdependence among constraints.

Case 1: Only focusing on line capacity safety.

Case 2: Focusing on both line capacity safety and voltage safety.

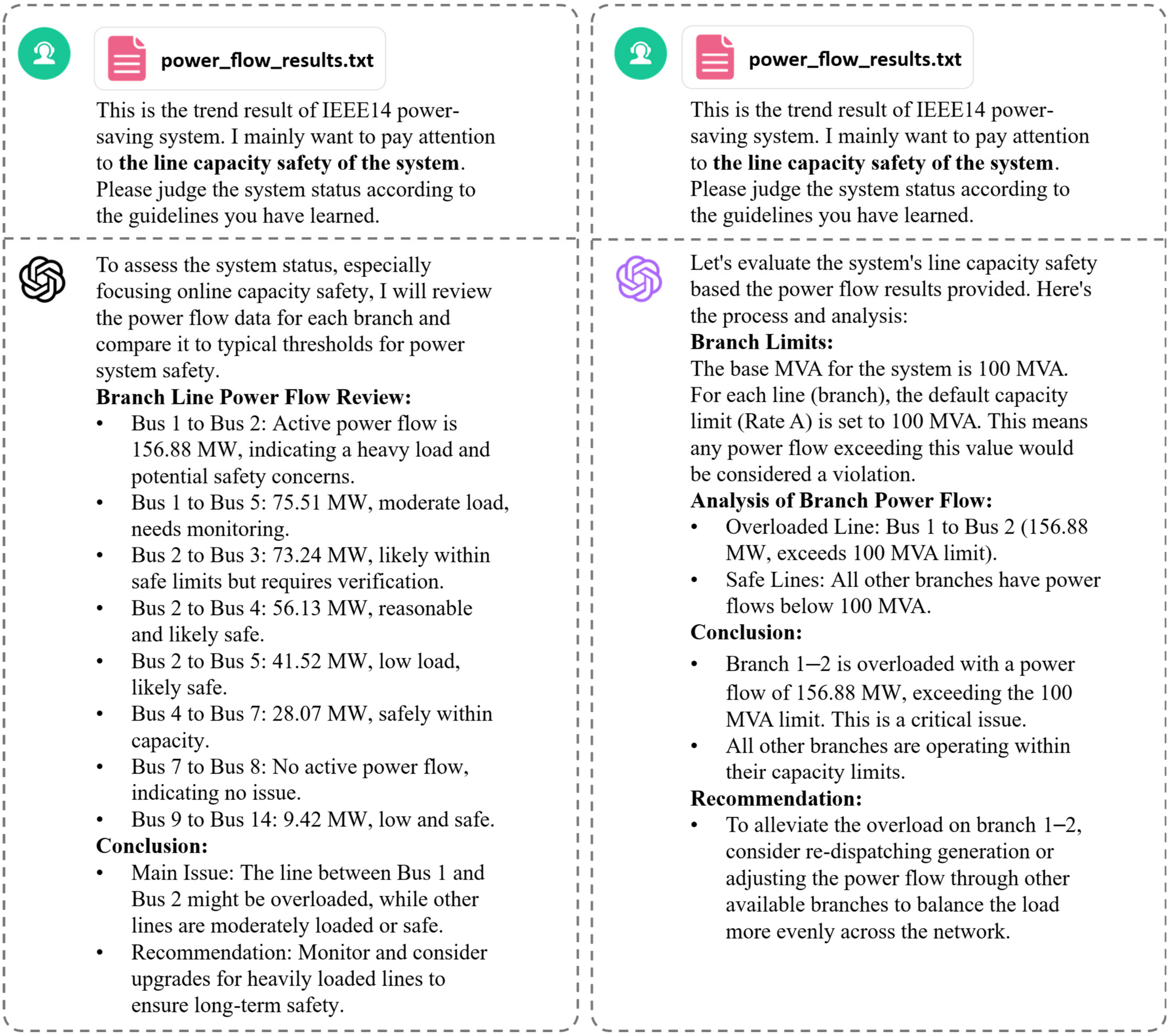
Case 3: Focusing on line capacity safety, voltage safety, and backup storage capacity safety.

The comparison between a general GPT-4 and a prompt- and document-conditioned GPT-4 was conducted to quantify the incremental benefit of injecting domain rules, threshold specifications, and structured output formats. Using the same base model controls for confounders related to architecture or capacity, thereby isolating improvements attributable solely to prompting and grounding. Results for Case 1 are shown in Figure 3.

In the experiment considering only line capacity safety, the conditioned variant outperformed the general model. As shown in the left panel, the general GPT-4—lacking explicit domain knowledge and task-specific prompts—tended to produce generic, hedged statements with weak binding to the thermal-limit criterion, leading to greater uncertainty and over-generalization and thus lower task effectiveness. By contrast, the right panel shows that the prompt- and document-conditioned model, informed by explicit rules and references, generated precise, severity-anchored outputs that were more specific and actionable. These findings indicate that equipping GPT-4 with domain-specific prompts and document grounding can substantially improve performance in specialized power-system monitoring tasks and enhance the practical value of its outputs.

Under the comprehensive fuzzy evaluation scheme, lower scores denote higher risk, with 0 indicating critical and 1 indicating safe. The general-purpose GPT-4 yielded a system safety score of 0.6, corresponding to a warning state, whereas the prompt- and document-conditioned variant produced 0.2, indicating a critical state. Given the pronounced overload on Line 12 relative to its 100 MVA rating, the correct classification is critical and immediate remedial action is warranted. The general model's underestimation arises from weak coupling between the numerical overload and the hard threshold, and from hedged, generic phrasing that biases template matching toward intermediate severity, which inflates the composite score. The conditioned model, explicitly anchored to domain rules, thresholds, and a structured output format, generates severity-anchored statements that align with the observed limit violation and drive the score into the critical region, thereby improving threshold fidelity and operational actionability while reducing the risk of delayed intervention. In addition, the model further provided simple dispatch recommendations, demonstrating its potential to go beyond passive detection and offer corrective guidance.

The test results for Scenario 2 are shown in Figure 4. In this test, the models considered both line capacity safety and voltage safety. Regarding line capacity issues, the general GPT-4 model provided relatively vague results, failing to identify specific safety issues in the system accurately. In contrast, the GPT-4, enhanced with prompt engineering and document learning, demonstrated stronger specificity, clearly identifying two specific issues: "Line 1-2 exceeding capacity limits". This improvement can be attributed to GPT-4's learning of relevant foundational knowledge, leading to a clearer definition of line capacity safety. In addition, the optimized model produced explicit adjustment suggestions, such as reducing generation at abnormal voltage nodes and load shedding at overloaded lines, which illustrates its capability to combine risk detection with actionable dispatch guidance.



power_flow_results.txt

This is the trend result of IEEE14 power-saving system. I mainly want to pay attention to **the line capacity safety of the system**. Please judge the system status according to the guidelines you have learned.

To assess the system status, especially focusing online capacity safety, I will review the power flow data for each branch and compare it to typical thresholds for power system safety.

Branch Line Power Flow Review:

- Bus 1 to Bus 2: Active power flow is 156.88 MW, indicating a heavy load and potential safety concerns.
- Bus 1 to Bus 5: 75.51 MW, moderate load, needs monitoring.
- Bus 2 to Bus 3: 73.24 MW, likely within safe limits but requires verification.
- Bus 2 to Bus 4: 56.13 MW, reasonable and likely safe.
- Bus 2 to Bus 5: 41.52 MW, low load, likely safe.
- Bus 4 to Bus 7: 28.07 MW, safely within capacity.
- Bus 7 to Bus 8: No active power flow, indicating no issue.
- Bus 9 to Bus 14: 9.42 MW, low and safe.

Conclusion:

- Main Issue: The line between Bus 1 and Bus 2 might be overloaded, while other lines are moderately loaded or safe.
- Recommendation: Monitor and consider upgrades for heavily loaded lines to ensure long-term safety.

power_flow_results.txt

This is the trend result of IEEE14 power-saving system. I mainly want to pay attention to **the line capacity safety of the system**. Please judge the system status according to the guidelines you have learned.

Let's evaluate the system's line capacity safety based the power flow results provided. Here's the process and analysis:

Branch Limits:

The base MVA for the system is 100 MVA. For each line (branch), the default capacity limit (Rate A) is set to 100 MVA. This means any power flow exceeding this value would be considered a violation.

Analysis of Branch Power Flow:

- Overloaded Line: Bus 1 to Bus 2 (156.88 MW, exceeds 100 MVA limit).
- Safe Lines: All other branches have power flows below 100 MVA.

Conclusion:

- Branch 1—2 is overloaded with a power flow of 156.88 MW, exceeding the 100 MVA limit. This is a critical issue.
- All other branches are operating within their capacity limits.

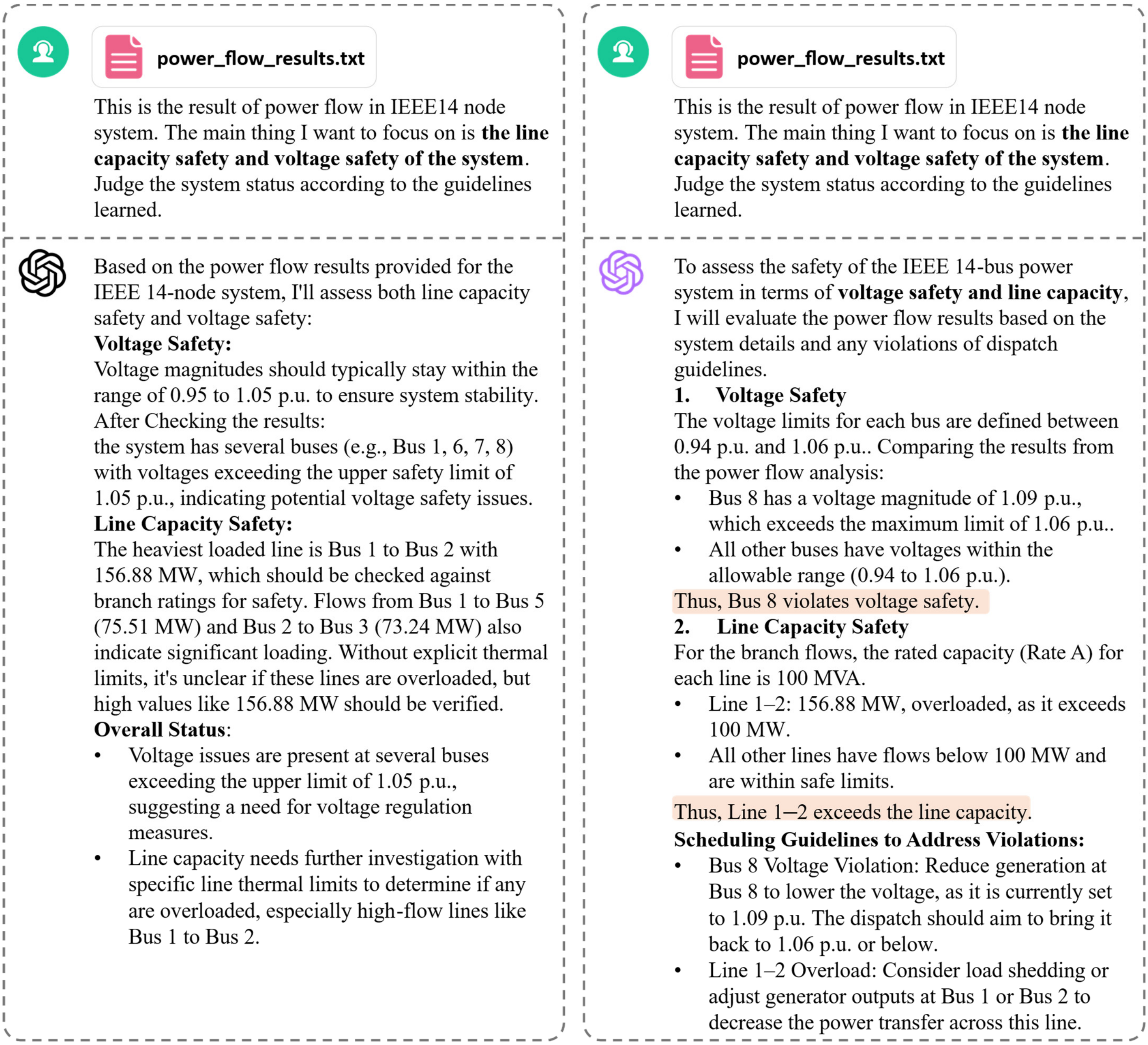
Recommendation:

- To alleviate the overload on branch 1—2, consider re-dispatching generation or adjusting the power flow through other available branches to balance the load more evenly across the network.

Figure 3. The test results only consider line capacity safety. The left image shows the test results using the general large model GPT-4, while the right image shows the test results using the GPT-4 model that was provided with specific prompts and relevant documents for learning.

The two models provided slightly different responses for voltage safety, particularly in their definition of voltage safety. The general GPT-4 defined the safe voltage range as 0.95 p.u. to 1.05 p.u., while the optimized GPT-4 used a range of 0.94 p.u. to 1.06 p.u., with the latter appearing more precise. However, upon further validation, it was found that the optimized GPT-4's answer was actually more accurate, as it correctly identified all voltage abnormal nodes, whereas the GPT-4 only highlighted the node with the most significant voltage issue. A possible reason for this discrepancy is that the GPT-4 optimized with prompts and relevant documents is based on the general GPT-4.

Using comprehensive fuzzy evaluation criteria for precise system state judgment, the general GPT-4 model yielded a comprehensive system safety score of 0.33, placing the system in a critical state. The prompt-optimized and document-learned GPT-4 produced a safety score of 0.2, indicating a critical state. In normal circumstances, the severe overload on Line 12 and the voltage limit violations at certain nodes would significantly impact the system's normal operation, warranting a critical state that requires immediate action. Both models agreed on the system's critical state, though their safety scores differed. The main reason for this discrepancy is that the general model lacks specialized knowledge in power system scheduling, leading to its suboptimal performance in system early warning.



power_flow_results.txt

This is the result of power flow in IEEE14 node system. The main thing I want to focus on is **the line capacity safety and voltage safety of the system**. Judge the system status according to the guidelines learned.

Based on the power flow results provided for the IEEE 14-node system, I'll assess both line capacity safety and voltage safety:

Voltage Safety:

Voltage magnitudes should typically stay within the range of 0.95 to 1.05 p.u. to ensure system stability.

After Checking the results:

the system has several buses (e.g., Bus 1, 6, 7, 8) with voltages exceeding the upper safety limit of 1.05 p.u., indicating potential voltage safety issues.

Line Capacity Safety:

The heaviest loaded line is Bus 1 to Bus 2 with 156.88 MW, which should be checked against branch ratings for safety. Flows from Bus 1 to Bus 5 (75.51 MW) and Bus 2 to Bus 3 (73.24 MW) also indicate significant loading. Without explicit thermal limits, it's unclear if these lines are overloaded, but high values like 156.88 MW should be verified.

Overall Status:

- Voltage issues are present at several buses exceeding the upper limit of 1.05 p.u., suggesting a need for voltage regulation measures.
- Line capacity needs further investigation with specific line thermal limits to determine if any are overloaded, especially high-flow lines like Bus 1 to Bus 2.

power_flow_results.txt

This is the result of power flow in IEEE14 node system. The main thing I want to focus on is **the line capacity safety and voltage safety of the system**. Judge the system status according to the guidelines learned.

To assess the safety of the IEEE 14-bus power system in terms of **voltage safety and line capacity**, I will evaluate the power flow results based on the system details and any violations of dispatch guidelines.

1. Voltage Safety

The voltage limits for each bus are defined between 0.94 p.u. and 1.06 p.u.. Comparing the results from the power flow analysis:

- Bus 8 has a voltage magnitude of 1.09 p.u., which exceeds the maximum limit of 1.06 p.u..
- All other buses have voltages within the allowable range (0.94 to 1.06 p.u.).

Thus, Bus 8 violates voltage safety.

2. Line Capacity Safety

For the branch flows, the rated capacity (Rate A) for each line is 100 MVA.

- Line 1–2: 156.88 MW, overloaded, as it exceeds 100 MW.
- All other lines have flows below 100 MW and are within safe limits.

Thus, Line 1—2 exceeds the line capacity.

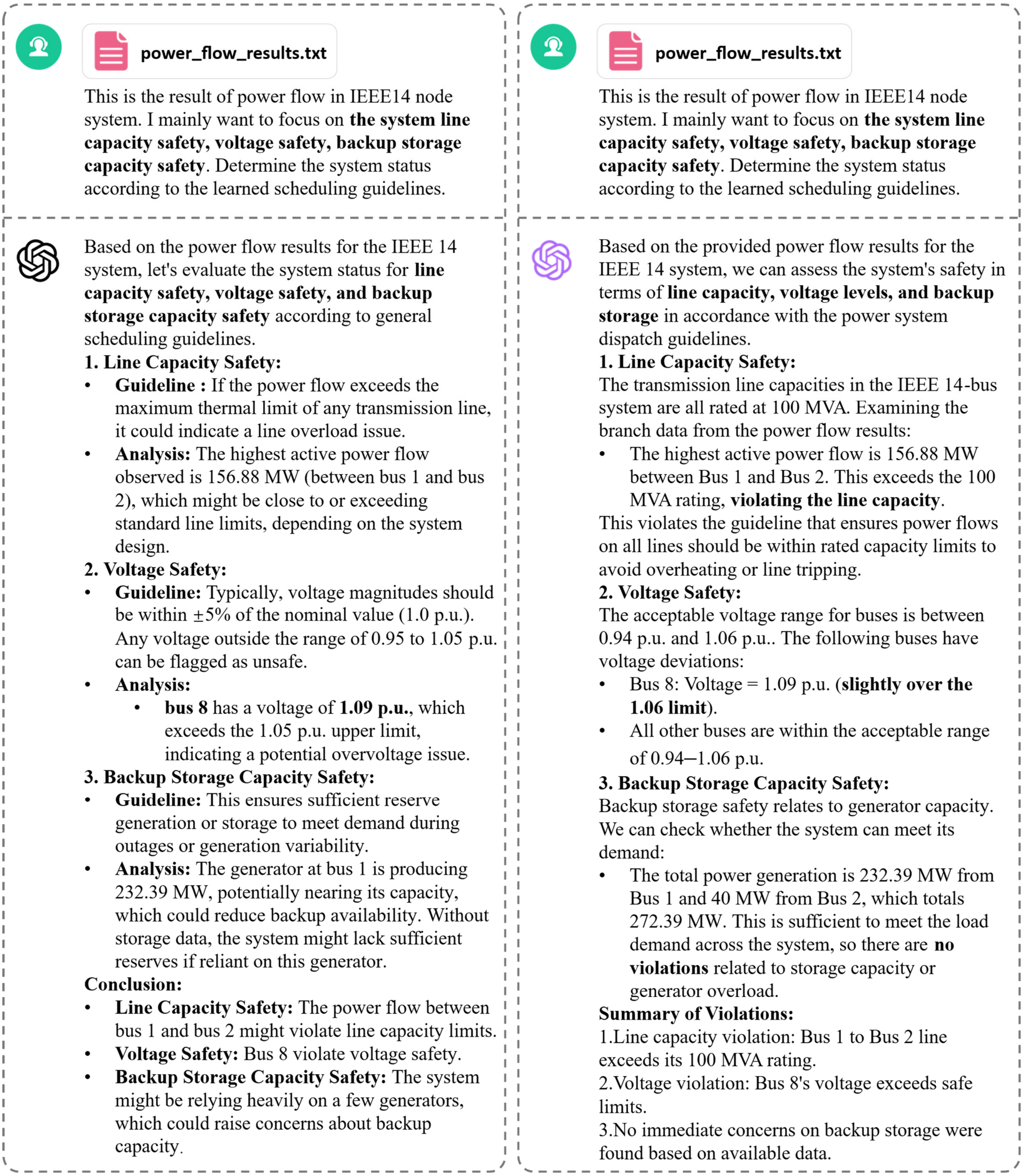
Scheduling Guidelines to Address Violations:

- Bus 8 Voltage Violation: Reduce generation at Bus 8 to lower the voltage, as it is currently set to 1.09 p.u. The dispatch should aim to bring it back to 1.06 p.u. or below.
- Line 1–2 Overload: Consider load shedding or adjust generator outputs at Bus 1 or Bus 2 to decrease the power transfer across this line.

Figure 4. The test results when focusing on both line capacity safety and voltage safety. The left image shows the test results using the general large model GPT-4, while the right image shows the test results using the GPT-4 model that was provided with specific prompts and relevant documents for learning.

Figure 5 reports the Case 3 results: both models reach consistent judgments on line capacity safety, as in earlier cases, indicating that thermal limit compliance is robust across configurations. The divergence arises in voltage safety. The general GPT-4 highlights only the violation at bus 8, prioritizing the most salient breach while overlooking the wider pattern of depressed voltages that typically accompanies congestion and depleted reactive margins. The optimized GPT-4 screens the full bus set against explicit thresholds, identifies all buses breaching or approaching limits, and relates these deviations to available reactive support and flow rerouting, thereby reproducing the diagnostic coverage established in Case 2. This broader assessment is operationally preferable because voltage restoration generally requires coordinated actions across multiple locations, including reactive dispatch, transformer tap adjustments, and redispatch, rather than a single-bus intervention. Consequently, the optimized configuration provides a completer and more consistent basis for remediation under multi-criteria monitoring, whereas the general GPT-4 shows a tendency to narrow attention as additional criteria are introduced.

Both models focused on meeting load demand for the backup capacity issue but provided slightly different responses. The general GPT-4 prioritized this issue due to the system's reliance on generators at nodes 1 and 2, while the optimized GPT-4 concluded the system had sufficient capacity. Upon validation, the optimized GPT-4 offered a more accurate response. Additionally, the optimized model demonstrated consistency and stability across cases, whereas the general GPT-4 tended to overlook certain issues when more safety criteria were introduced, leading to variations in its responses.



power_flow_results.txt

This is the result of power flow in IEEE14 node system. I mainly want to focus on **the system line capacity safety, voltage safety, backup storage capacity safety**. Determine the system status according to the learned scheduling guidelines.

Based on the power flow results for the IEEE 14 system, let's evaluate the system status for **line capacity safety, voltage safety, and backup storage capacity safety** according to general scheduling guidelines.

1. Line Capacity Safety:

- **Guideline :** If the power flow exceeds the maximum thermal limit of any transmission line, it could indicate a line overload issue.
- **Analysis:** The highest active power flow observed is 156.88 MW (between bus 1 and bus 2), which might be close to or exceeding standard line limits, depending on the system design.

2. Voltage Safety:

- **Guideline:** Typically, voltage magnitudes should be within ±5% of the nominal value (1.0 p.u.). Any voltage outside the range of 0.95 to 1.05 p.u. can be flagged as unsafe.
- **Analysis:**
    - **bus 8** has a voltage of **1.09 p.u.**, which exceeds the 1.05 p.u. upper limit, indicating a potential overvoltage issue.

3. Backup Storage Capacity Safety:

- **Guideline:** This ensures sufficient reserve generation or storage to meet demand during outages or generation variability.
- **Analysis:** The generator at bus 1 is producing 232.39 MW, potentially nearing its capacity, which could reduce backup availability. Without storage data, the system might lack sufficient reserves if reliant on this generator.

Conclusion:

- **Line Capacity Safety:** The power flow between bus 1 and bus 2 might violate line capacity limits.
- **Voltage Safety:** Bus 8 violate voltage safety.
- **Backup Storage Capacity Safety:** The system might be relying heavily on a few generators, which could raise concerns about backup capacity.

power_flow_results.txt

This is the result of power flow in IEEE14 node system. I mainly want to focus on **the system line capacity safety, voltage safety, backup storage capacity safety**. Determine the system status according to the learned scheduling guidelines.

Based on the provided power flow results for the IEEE 14 system, we can assess the system's safety in terms of **line capacity, voltage levels, and backup storage** in accordance with the power system dispatch guidelines.

1. Line Capacity Safety:

The transmission line capacities in the IEEE 14-bus system are all rated at 100 MVA. Examining the branch data from the power flow results:

- The highest active power flow is 156.88 MW between Bus 1 and Bus 2. This exceeds the 100 MVA rating, **violating the line capacity**.

This violates the guideline that ensures power flows on all lines should be within rated capacity limits to avoid overheating or line tripping.

2. Voltage Safety:

The acceptable voltage range for buses is between 0.94 p.u. and 1.06 p.u.. The following buses have voltage deviations:

- Bus 8: Voltage = 1.09 p.u. (**slightly over the 1.06 limit**).
- All other buses are within the acceptable range of 0.94—1.06 p.u.

3. Backup Storage Capacity Safety:

Backup storage safety relates to generator capacity. We can check whether the system can meet its demand:

- The total power generation is 232.39 MW from Bus 1 and 40 MW from Bus 2, which totals 272.39 MW. This is sufficient to meet the load demand across the system, so there are **no violations** related to storage capacity or generator overload.

Summary of Violations:

1.Line capacity violation: Bus 1 to Bus 2 line exceeds its 100 MVA rating.

2.Voltage violation: Bus 8's voltage exceeds safe limits.

3.No immediate concerns on backup storage were found based on available data.

Figure 5. The test results when focusing on line capacity safety, voltage safety, and backup capacity safety simultaneously. The left image shows the test results using the general large model GPT-4, while the right image shows the test results using the GPT-4 model that was provided with specific prompts and relevant documents for learning.

Using comprehensive fuzzy evaluation criteria for precise system state judgment, the general GPT-4 gave a system safety score of 0.46, indicating that the system was in an emergency state. The optimized GPT-4 provided a system safety score of 0.54, indicating an emergency state. In normal dispatch scenarios, since the system had sufficient backup capacity, adjusting the output of system generators could resolve both the line congestion and voltage limit safety issues, making the emergency state result reasonable. The slightly higher score from the optimized model is consistent with its recognition of viable remedial actions through generator reallocation and voltage support, yielding an accurate and operationally actionable judgment.

To avoid randomness in the model evaluation in any given scenario, we calculated the scores of the two models across the three scenarios for each metric. By averaging these scores, we obtained the results. The test outcomes are as shown in Table 2. The detailed metric values for the three scenarios are provided in Appendix A (Table A1).

Table 2. Evaluation on Power System Monitoring and Safety Assessment.

Model	F_a	Log	S_t	S_{SMOG}	R_e	D_{key}
GPT-4	0.95	0.83	1	0.66	0.15	0.71
GPT-4 after learning	0.95	0.88	1	0.14	0.21	0.76

Figure 6 shows that both models perform similarly in terms of authenticity, stability, and information redundancy, with acceptable and reliable results. However, the GPT-4 model trained with prompts and document learning outperforms in logic, key information density, and SMOG readability, demonstrating better language structure, higher readability, and more concise information.

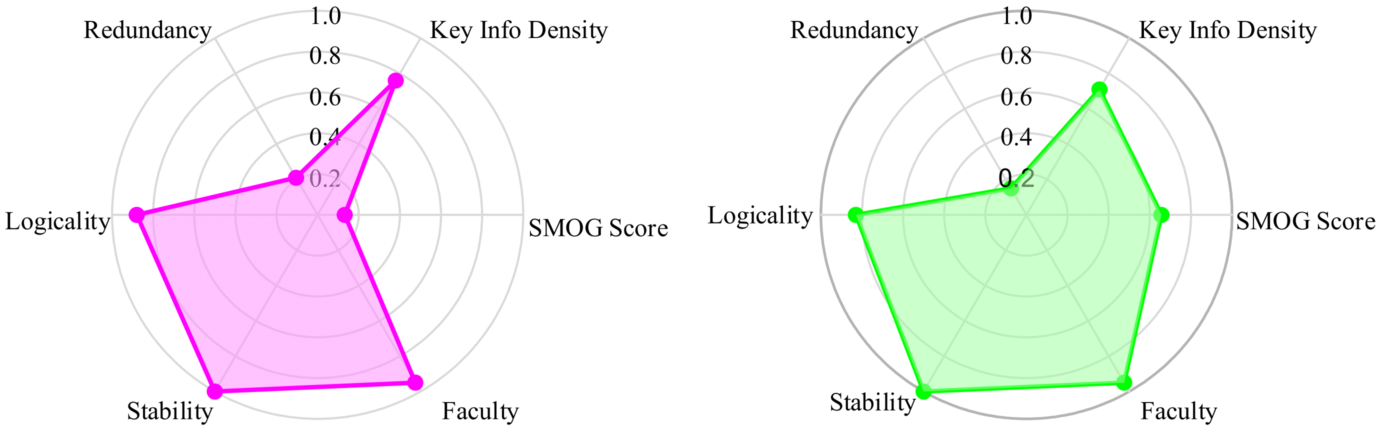


Figure 6. The test results for the metrics. The (**left**) image represents the results from testing on the general large model, GPT-4, while the (**right**) image displays the test results on the GPT-4 model that received a prompt and related document learning.

4. Discussion and Conclusions

This paper has proposed an intelligent early warning method for secure operation of power systems based on LLMs, which combines domain-specific prompt engineering with fuzzy evaluation to achieve real-time state assessment and generate dispatch recommendations. The proposed framework establishes a closed-loop process, where potential security violations are not only detected and quantified but also accompanied by targeted remedial actions, thereby linking monitoring, assessment, and response. Testing on the IEEE 14-bus system has demonstrated the feasibility and effectiveness of the method. The prompt-optimized model exhibited superior performance in terms of logical consistency, stability, and key information density compared with a general-purpose LLM. Its early warning results and corresponding dispatch suggestions proved more accurate and actionable, thus improving operational efficiency and closing the loop between risk detection and corrective decision support.

However, despite the positive results, some limitations remain. First, the model requires further customization to learn more specialized knowledge of power systems, enhancing its sensitivity to power flow results and system parameters. Second, the transparency and interpretability of the model's decisions need improvement to increase dispatchers' trust in the model's output. Additionally, the model's adaptability to different grid structures needs enhancement, especially when encountering unfamiliar grid topologies or emergencies, where its performance may degrade. Future research can be pursued in the following directions:

- Domain-specific adaptation and interpretability: Further fine-tuning with power system data and incorporating expert rules or explainable AI methods will improve both the sensitivity of the framework to grid dynamics and the transparency of its decision-making, thereby increasing operator trust [3,8,24].
- Hybrid intelligence with advanced optimization: Combining LLMs with complementary AI techniques such as reinforcement learning, knowledge graphs, and graph-embedding approaches including PowerGraph-LLM [37], can enhance robustness, scalability, and optimization capability in complex grid environments [38,39].
- Extension to multi-energy systems under uncertainty: Applying the proposed framework to integrated energy hubs, particularly under uncertain operating conditions as studied by Giannelos et al. [40], would broaden its applicability and support risk-aware multi-energy management [31].

Appendix A

Table A1. Quantitative evaluation metrics of LLMs under three test scenarios.

Case	Model	F_a	Log	S_t	S_{SMOG}	R_e	D_{key}
Case 1	GPT-4	0.91	0.39	0.99	12.40	0	0.25
	GPT-4 after learning	0.96	0.46	0.99	12.34	0	0.30
Case 2	GPT-4	0.98	0.27	0.99	15.08	0.17	0.22
	GPT-4 after learning	0.94	0.30	1.00	11.21	0.30	0.21
Case 3	GPT-4	0.95	0.25	1.00	13.78	0.13	0.27
	GPT-4 after learning	0.96	0.26	1.00	11.62	0.12	0.25

Acknowledgments

The authors would like to thank all those who provided support and constructive feedback during the preparation of this work.

Author Contributions

Conceptualization, Y.X.; Methodology, L.T. and Y.X.; Writing—Original Draft Preparation, L.T.; Writing—Review & Editing, L.T., Y.X., G.Q., Z.T. and J.L.; Visualization, L.T.; Project Administration, Y.X.; Funding Acquisition, Y.X.

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data and materials supporting the findings—prompts, rule templates, topology descriptions, scenario configurations, raw numerical tables for Figures 3–5, fuzzy-evaluation parameters, and the replication protocol—are included in the manuscript's Materials and Methods section and Appendix.

Funding

This research was funded by the Sichuan University Graduate Course Construction Project "Power Dispatch Intelligent Brain—Large Language Model Based Power Security Early Warning and Economic Dispatch Practice Teaching Platform" (2024CJAL004), and the Sichuan University "Artificial Intelligence Empowerment for Innovative Practice Education Reform" project entitled "Application and Education of Artificial Intelligence Large Language Models in Power System Operation and Control".

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Sun H, Huang T, Guo Q, Zhang M, Guo W, Liu W, et al. Research and application of an intelligent machine dispatcher for dispatch decision-making. *Power Syst. Technol.* **2020**, *44*, 1–8. doi:10.13335/j.1000-3673.pst.2019.1937.
2. Shi G, Qiu X, Zhao J, Ma J. Rolling dispatch of power systems considering wind power prediction errors and demand response. *Mod. Electr. Power* **2018**, *35*, 9–15. doi:10.19725/j.cnki.1007-2322.2018.06.002.
3. Zhang J, Xu J, Xu P, Chen S, Gao T, Bai Y. Review and prospects of the application of large AI models in power system operation control. *J. Wuhan Univ. (Eng. Ed.)* **2023**, *56*, 1368–1379. doi:10.14188/j.1671-8844.2023-11-008.
4. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
5. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
6. OpenAI. ChatGPT. Available online: http://chat.openai.com/ (accessed on 28 August 2025).
7. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. doi:10.1145/3560815.
8. Cheng Y, Zhao H, Zhou X, Zhao J, Cao Y, Yang C. GAIA: A large language model for advanced power dispatch. *arXiv* **2024**, arXiv:2408.03847.
9. Lai Z, Wu T, Fei X, Ling Q. BERT4ST: Fine-tuning pre-trained large language model for wind power forecasting. *Energy Convers. Manag.* **2024**, *307*, 118331. doi:10.1016/j.enconman.2024.118331.
10. Gao M, Zhou S, Gu W, Wu Z, Liu H, Zhou A. A general framework for load forecasting based on pre-trained large language model. *arXiv* **2024**, arXiv:2406.11336.
11. Liang X, Zhang W, Lei S, Zhang Y, Xu M, Peng L, et al. Multi-classification of electric power metadata based on prompt-tuning. In *Artificial Intelligence and Mobile Services—AIMS 2022*; Pan X, Jin T, Zhang L-J, Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 102–114. doi:10.1007/978-3-031-23504-7_8.
12. Cheng Y, Zhao H, Zhou X, Zhao J, Cao Y, Yang C, et al. A large language model for advanced power dispatch. *Sci. Rep.* **2025**, *15*, 8925. doi:10.1038/s41598-025-91940-x.
13. Jiang G, Ma ZH, Zhang L, Chen J. EPlus-LLM: A large language model-based computing platform for automated building energy modeling. *Appl. Energy* **2024**, *367*, 123431. doi:10.1016/j.apenergy.2024.123431.
14. Liu F, Tong X, Yuan M, Zhang Q. Algorithm evolution using large language model. *arXiv* **2023**, arXiv:2311.15249.
15. Li B, Jiang Y, Xu J, Liu Z, Sheng Z, Song X, et al. The design and implementation of an intelligent Q&A system for electric power safety regulations based on large language model technology. In Proceedings of the 2024 6th International Conference on Energy Systems and Electrical Power (ICESEP), Wuhan, China, 2024; pp. 604–607. doi:10.1109/ICESEP62218.2024.10652122.
16. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
17. National Technical Committee 446 on Grid Operation and Control of Standardization Administration of China (SAC/TC 446). *Grid Operation Code (GB/T 31464-2022)*; China Standard Press: Beijing, China, 2022.
18. Glover JD, Sarma MS, Overbye TJ. *Power System Analysis & Design*, 6th ed.; Cengage Learning: Boston, MA, USA, 2017.
19. Zimmerman RD, Murillo-Sánchez CE, Thomas RJ. MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education. *IEEE Trans. Power Syst.* **2011**, *26*, 12–19. doi:10.1109/TPWRS.2010.2051168.
20. International Electrotechnical Commission (IEC). *IEC 61970-301: Energy Management System Application Program Interface (EMS-API)—Part 301: Common Information Model (CIM) Base*; IEC: Geneva, Switzerland, 2020.
21. Blecher L, Cucurull G, Scialom T, Stojnic R. Nougat: Neural optical understanding for academic documents. *arXiv* **2023**, arXiv:2308.13418.
22. Wei J, Wang X, Schuurmans D, Stojnic R. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
23. Yan Z, Xu Y. Real-time optimal power flow with linguistic stipulations: integrating GPT-agent and deep reinforcement learning. *IEEE Trans. Power Syst.* **2024**, *39*, 4747–4750. doi:10.1109/TPWRS.2023.3338961.
24. State Council of the People's Republic of China. *Regulations on the Emergency Disposal, Investigation and Handling of Electric Power Safety Accidents (State Council Decree No. 599)*; State Council: Beijing, China, 2011.
25. Reimers J, Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 2019; pp. 3982–3992. doi:10.48550/arXiv.1908.10084.
26. Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 2021; pp. 6894–6910. doi:10.48550/arXiv.2104.08821.
27. Takagi T, Sugeno M. Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Trans. Syst. Man Cybern.* **1985**, *SMC-15*, 116–132. doi:10.1109/TSMC.1985.6313399.

28. Zimmermann H-J. *Fuzzy Set Theory—and Its Applications*, 4th ed.; Springer Science & Business Media: Berlin, Germany, 2010.
29. Zadeh LA. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. doi:10.1016/S0019-9958(65)90241-X.
30. Huang Z, Shi G, Sukhatme GS. From words to routes: Applying large language models to vehicle routing. *arXiv* **2024**, arXiv:2403.10795.
31. Muhammad Y, Khan R, Raja MAZ, Ullah F, Chaudhary NI, He Y. Solution of optimal reactive power dispatch with FACTS devices: A survey. *Energy Rep.* **2020**, *6*, 2211–2229. doi:10.1016/j.egyr.2020.07.030.
32. Luo H, Sun Q, Xu C, Zhao P, Lou J, Tao C, et al. WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv* **2023**, arXiv:2308.09583.
33. Jin Z, Chen Y, Leeb F, Gresele L, Kamal O, Lyu Z, et al. CLadder: A benchmark to assess causal reasoning capabilities of language models. *arXiv* **2023**, arXiv:2312.04350.
34. Wang B, Xu C, Wang S, Gan Z, Cheng Y, Gao J, et al. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *arXiv* **2021**, arXiv:2111.02840.
35. Ye J, Wu Y, Gao S, Huang C, Li S, Li G, et al. RoTBench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. *arXiv* **2024**, arXiv:2401.08326.
36. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. doi:10.1016/j.lindif.2023.102274.
37. Bernier F, Cao J, Cordy M, Ghamizi S. PowerGraph-LLM: Novel Power Grid Graph Embedding and Optimization with Large Language Models. *IEEE Trans. Power Syst.* **2025**, *early access*. doi:10.1109/TPWRS.2025.3596774.
38. Bilal A, Ebert D, Lin B. LLMs for Explainable AI: A Comprehensive Survey. *arXiv* **2025**, arXiv:2504.00125.
39. Zhou X, Zhao H, Cheng Y, Cao Y, Liang G, Liu G, et al. ElecBench: A Power Dispatch Evaluation Benchmark for Large Language Models. *arXiv* **2024**, arXiv:2407.05365.
40. Giannelos S, Pudjianto D, Zhang T, Strbac G. Energy Hub Operation Under Uncertainty: Monte Carlo Risk Assessment Using Gaussian and KDE-Based Data. *Energies* **2025**, *18*, 1712. doi:10.3390/en18071712.