scientific reports

Check for updates

OPEN

A large language model for advanced power dispatch

Yuheng Cheng[1,2,8], Huan Zhao[3,8], Xiyuan Zhou[4], Junhua Zhao[1,2]✉, Yuji Cao[5], Chao Yang[6] & Xinlei Cai[7]

Power dispatch is essential for providing society with stable, cost-effective, and eco-friendly electricity. However, traditional methods falter as power systems grow in scale and complexity, struggling with multitasking, swift problem-solving, and human-machine collaboration. This paper introduces Grid Artificial Intelligent Assistant (GAIA), a pioneering Large Language Model (LLM) designed to assist with a variety of power system operational tasks, including operation adjustment, operation monitoring, and black start scenarios. We have developed a novel dataset construction technique that harnesses various data sources to fine-tune GAIA for optimal performance in this domain. This approach streamlines LLM training, allowing for the seamless integration of multidimensional data in power system management. Additionally, we have crafted specialized prompt strategies to boost GAIA's input-output efficiency in dispatch scenarios. When evaluated on the ElecBench benchmark, GAIA surpasses the baseline model Large Language Model Meta AI-2 (LLaMA2) on multiple metrics. In practical applications, GAIA has demonstrated its ability to enhance decision-making processes, improve operational efficiency, and facilitate better human-machine interactions in power dispatch operations. This paper expands the application of LLMs to power dispatch and validates their practical utility, paving the way for future innovations in this field.

Ensuring power system stability and economic efficiency hinges on safe and effective power dispatch[1]. System operators must skillfully balance generating unit outputs and load distribution across transmission lines, adapting to dynamic power supply and demand shifts due to human activities, weather variations, and emergencies. Some of these operational tasks are directly related to the decision process, such as Economic Dispatch (ED) and Unit Commitment (UC) problems, while others support the decision-making or execution of power dispatch by human dispatchers. For example, tasks like automated Question and Answer systems can be designed to extract decision-related information, and operation monitoring ensures the safe execution of operational plans. Dispatch processes must account for demand changes, generation unit costs and conditions, and transmission line capacities, all while meeting the power market's economic and reliability standards. The growing integration of renewable energy resources and advancements in HVDC technology further complicate dispatch operations, demanding more advanced methods[2].

Traditional optimization algorithms, such as linear and nonlinear programming[3], are effective for specific, well-defined power dispatch problems. However, recent advancements like[4] focusing on renewable integration still rely on deterministic models that inadequately capture operational uncertainties. At the same time,[5] framework, despite improving convergence, remains impractical for real-time dispatch due to excessive computational latency. These methods often struggle with the increasing complexity introduced by renewable energy volatility and load fluctuations. In contrast, deep learning and reinforcement learning[6] offer data-driven adaptability but face deployment barriers:[7] black-box load forecasting models hinder operator trust in critical scenarios, and[8] multi-agent system lacks natural language interfaces for collaborative decision-making. This creates a critical gap in solutions that simultaneously ensure computational efficiency, scenario adaptability, and human-centric interaction - a gap addressed by GAIA through its novel fusion of domain-specific knowledge engineering and LLM-based natural language processing.

Recent breakthroughs in Large Language Models (LLMs)[9] have revolutionized their capabilities. Models like Transformer, LLaMA[10], and ChatGPT[11] have achieved mastery over language's deep structures and context

[1]Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen 518129, China. [2]Chinese University of Hong Kong, School of Science and Engineering, Shenzhen 518172, China. [3]Department of Building Environment and Energy Engineering, Hong Kong Polytechnic University, Hong Kong, China. [4]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore. [5]Department of Mechanical and Automation Engineering, Chinese University of Hong Kong, Hong Kong 999077, China. [6]School of Electrical and Electronic Engineering, North China Electric Power University, Baoding 071003, China. [7]China Southern Power Grid (China), Guangzhou 510600, China. [8]Yuheng Cheng and Huan Zhao contributed equally to this work. ✉email: zhaojunhua@cuhk.edu.cn

through extensive pre-training, allowing them to comprehend and follow complex instructions. These LLMs can perform specific tasks at or above human levels when finely tuned and given detailed instructions. Prompt engineering[12] further enhances their adaptability, enabling them to tackle new tasks without needing large-scale retraining. This flexibility significantly benefits complex decision-making and human-machine interaction. LLMs' robust natural language processing skills are essential for efficient collaboration between humans and machines. Their strong generalization ability also minimizes the necessity for new models for different scenarios, showcasing their versatility and potential for widespread application.

Despite the success of LLMs in different domains, there has yet to be a dedicated LLM for power dispatch. Existing general-purpose LLMs like GPT-4[13] or fine-tuned LLMs for the mathematics field like WizardMath[14] cannot adequately address the problems in power dispatch. The main barriers to building power dispatch LLM are listed as follows:

1. **Lack of domain dataset**: The data in power systems is multiple perspective[15], such as load, cost, topology, etc. LLMs need to integrate these data to understand and learn the characteristics of power systems. The training data for other LLMs mostly consists of textual or basic numerical reasoning data, lacking actual power grid operation information.
2. **Specific domain adaption**: LLMs for power dispatch need to be tailored to grasp the power sector's unique jargon and decision-making processes. This requires not only proficiency in specific terminology but also a comprehensive understanding of industry-specific scenarios, such as load forecasting and unit commitment, which are not inherent in general-purpose LLMs.
3. **Complex human-machine interaction**: The prompts for LLM in power dispatch scenarios significantly influence the performance of results. The prompts should be closely integrated with the power dispatch operations, containing key dispatching terminologies and conforming to related operational rules and decision-making processes. So that dispatchers can interact with LLM naturally using natural language and understand and trust LLM's results.

To overcome the above challenges, this paper seeks to enhance intelligent power dispatch and human-machine collaboration by advancing the use of LLMs in the energy sector. Our key contributions are:

1. We develop GAIA, a model that assists dispatchers by providing operational suggestions in scenarios like operation adjustment, operation monitoring, and black start.
2. We introduce a dedicated pipeline for data generation, processing, and LLM training tailored to power dispatch challenges.
3. We categorize power dispatch scenarios and design specific prompts for targeted training, incorporating techniques for interacting with LLMs, such as vector data representation, text data enrichment, and specialized terminology handling.

Our evaluation of GAIA in the ElecBench benchmark[16] demonstrates its superiority over baseline model LLaMA2[17], across various dispatch-related metrics. The remaining sections of this paper are as follows: Section "Domain-specific large language models" introduces the related works of domain-specific LLMs. The pipeline for GAIA is proposed in Sect. "Methodology", including the division of dispatch scenarios, data generation and collection methods, and training techniques. Section "Performance analysis for power-related tasks" presents the model validation results, followed by Sect. "Conclusion", which is the discussion and conclusion.

Domain-specific large language models

Domain-specific LLMs have achieved significant progress in recent years. These advancements are largely attributed to the development of innovative dataset generation methods, efficient parameter fine-tuning techniques, and model customization for specific domains. The following contents summarize representative research in these key areas, collectively contributing to developing domain-specific LLMs.

Specific large language models

Domain-specific LLMs have shown impressive capabilities in fields like mathematics, computer science, healthcare, and power systems. These models are refined from general LLMs using domain-specific data to meet particular needs.

For example, in the field of mathematics, WizardMath[14] is fine-tuned from the LLaMA2 model to enhance the mathematical reasoning ability, by generating diverse mathematical instruction data through Evol-Instruct. In computer programming tasks, CodeLLaMA[18] uses the Self-Instruct method to generate datasets in the LLaMA2 70b model, which is then fine-tuned after validation. In the medical field, Med-PaLM[19] utilized Instruction Tuning[20] to fine-tune the Flan-PaLM[21] model efficiently, and its training data is obtained through random selection and manual evaluation filtering. For short-term load forecasting tasks in power systems, LFLLM[22] presents an efficient training method based on PEFT to address the challenging training problem of LLMs with massive parameters, ensuring excellent learning capability of the model.[23] introduced a comprehensive framework for leveraging LLMs in power systems, demonstrating their potential in tasks such as fault diagnosis and state estimation.[24] provided a detailed analysis of the strengths and weaknesses of LLMs in the electric energy sector, highlighting areas where they excel and where further research is needed.

The effectiveness of domain-specific LLMs stems from their strategic use and creation of enhanced datasets, along with advanced fine-tuning techniques, allowing them to tackle complex domain-specific reasoning tasks. Despite their proven potential across various sectors, there's a gap in the literature concerning data. While

these models are adapted to particular applications using specialized domain datasets, there's a need for more systematic approaches in dataset generation, selection, optimization, and fine-tuning methods.

Dataset generation methods

Constructing a domain-specific LLM requires a series of steps to produce a relevant dataset. This involves selecting training tasks, analyzing and generating data, refining the dataset, and thorough filtering and post-processing to improve quality and diversity. The initial step is identifying the necessary training tasks for the domain. For instance, DARWIN[25], a natural science LLM, categorizes its training tasks into material inverse design, property classification, and attribute regression prediction. During the data collection phase, domain-related papers, books, and news are typically converted into a processable text format. Subsequently, in the auto-generation of fine-tuning datasets phase, LLMs are often utilized to generate question-answer pairs. Among these, Self-Instruct[26] enhances the model's ability to understand instructions by generating instructions by itself. Evol-Instruct[27] uses evolutionary algorithms to create complex code instructions, improving the model's fine-tuning performance on code generation tasks. Explore-Instruct[28] employs an active learning strategy to explore and expand the specific domain's set of instructions, optimizing the model's task-processing performance. In the filtering and post-processing stage, irrelevant or low-quality data from the generated results are typically removed.

Existing research on domain-specific LLMs covers task selection, data collection, dataset generation, and post-processing. However, it frequently overlooks the unique needs of intricate fields like power system dispatch. These include handling specialized terms, complex formulas, and simulation data. Additionally, current methods often neglect to develop complex problem-solving skills from simulation data in dataset generation, a critical aspect for applications in power system dispatch.

Baseline LLM

Many open-source LLMs are available, each with unique strengths. MPT[29] excels in reasoning over long texts and offers optimized training and inference speeds. Falcon[30] enhances performance by selectively processing network data for pre-training datasets. Meta's LLaMA2, an improved version of LLaMA, demonstrates strong logical and mathematical reasoning capabilities. In benchmarks like MMLU[31] and MATH[32], LLaMA2 outperforms competitors with similar model sizes.

LLaMA2 is an LLM based on the Transformer Decoder architecture, which has been optimized on top of the standard structure, such as introducing RMSNorm[33] pre-normalization layer, SwiGLU[34] activation function, and Rotary Position Embedding (RoPE)[35]. The training dataset for LLaMA2 is massive, reaching 20 trillion tokens, which is a 40% increase compared to the previous generation model, and the context length has been extended from 2048 to 4096 tokens. This allows the model to understand and generate longer texts.

A notable feature of the LLaMA2 is the adoption of the Grouped-Query Attention[36] mechanism. This optimized attention mechanism improves inference throughput by reducing the size of the key-value cache. The Grouped-Query Attention reduces memory usage by sharing a single key-value projection, and experiments have shown that it performs comparably to Multi-Head Attention[37] on most evaluation tasks and typically outperforms Multi-Query Attention[38].

To enhance the model's performance, LLaMA2 also uses the Reinforcement Learning from Human Feedback[39] method, which iteratively fine-tunes the model with human feedback data, including using rejection sampling and Proximal Policy Optimization[40].

In summary, the LLaMA2 excels in various benchmark tests due to its sophisticated architecture and training techniques. Its versatility and precision in generating responses establish it as a significant benchmark in the LLM field.

Parameter efficient fine-tuning methods

The Parameter-Efficient Fine-Tuning (PEFT)[41] is a class of fine-tuning methods for computational resource-limited situations. Selecting the suitable PEFT method is crucial for improving the LLM performance of downstream tasks. Adapter[42] achieves parameter-efficient task-specific adjustments by inserting small network layers into pre-trained models. Still, this method may introduce additional computational overhead and may be less effective than full model fine-tuning in some tasks. P-Tuning[43] fine-tunes models by adding learnable continuous vectors (i.e., prompts). Although this method has advantages in parameter efficiency, it relies on the model's sensitivity to prompts and may require larger training datasets to optimize prompts.

To address these limitations, Low-Rank Adaptation (LoRA)[44] reduces the number of fine-tuning parameters and alleviates computational burden by performing low-rank decomposition on weight matrices and updating small matrices. Further, Quantized Low-Rank Adaptation (QLoRA)[45] decreases the representation precision of low-rank matrices using quantization techniques, which reduces not only the number of parameters but also the model's memory footprint and computational complexity, making the fine-tuning process more efficient in resource-constrained environments. These methods demonstrate the ability to adapt to specific tasks and to preserve pre-trained model knowledge by fine-tuning a few parameters, highlighting the importance of choosing suitable fine-tuning strategies in different scenarios.

Methodology

This section outlines a robust methodology for creating an LLM tailored to power dispatch tasks. Our goal is to build a model that comprehends power system dynamics, facilitates human-computer interaction, and delivers precise decisions. We employ a multi-stage pipeline-comprising data generation, preprocessing, training, and optimization leveraging multi-task learning and targeted training to refine the LLM's capabilities in the power

sector. We detail the training data composition and methods for processing text and simulation data, offering theoretical and practical foundations for the model's development.

Framework

The objective of building an LLM for power dispatch is clear: the LLM can grasp power system dynamics, aid dispatchers through human-machine collaboration, and consistently make precise decisions. Current LLMs struggle with the power industry's unique challenges, such as interpreting diverse power system data, specialized jargon, and real-world dispatch scenarios.

Addressing these issues, We hypothesize that a specially trained LLM can enhance power dispatch operations. Utilizing a robust dataset, multi-task learning, and specialized fine-tuning, the LLM can achieve higher predictive accuracy and offer improved decision support for dispatchers. We've devised a multi-stage pipeline: data generation and preprocessing, interactive prompt design, and model training and optimization. This approach seeks to surpass traditional power dispatch limitations and bolster intelligent power system operations.

As shown in Fig. 1, the overall framework's pipeline includes the following three key steps:

1. **Simulation data processing**: We select simulation scenarios that align with dispatch operations: adjustment, monitoring, and black start. Each scenario has a dedicated simulation program to produce system data. This data is then transformed into structured Q&A pairs using templates, enhancing the model's comprehension and learning.
2. **Text data processing**: We employ Optical Character Recognition (OCR) to digitize power-related documents, converting them to editable text. This text is then segmented and paired with prompts. Finally, GPT-4 generates additional Q&A pairs, enriching the training dataset.
3. **Training of GAIA**: The LLaMA2 undergoes supervised fine-tuning with extensive Q&A data from simulations and texts. This targeted training sharpens the model's performance on power dispatch tasks, ensuring greater efficiency and precision in real-world operations.

A crucial aspect of GAIA's design is the iterative feedback loop between GAIA and human dispatchers. After GAIA generates responses or suggestions, dispatchers provide feedback, indicating their agreement or disagreement. Crucially, GAIA is designed as a decision support tool and does not have direct control over power system equipment. It provides recommendations to dispatchers, who retain ultimate control and decision-making authority. While GAIA can access real-time information from the Energy Management System (EMS) to inform its suggestions, it does not execute any control actions.

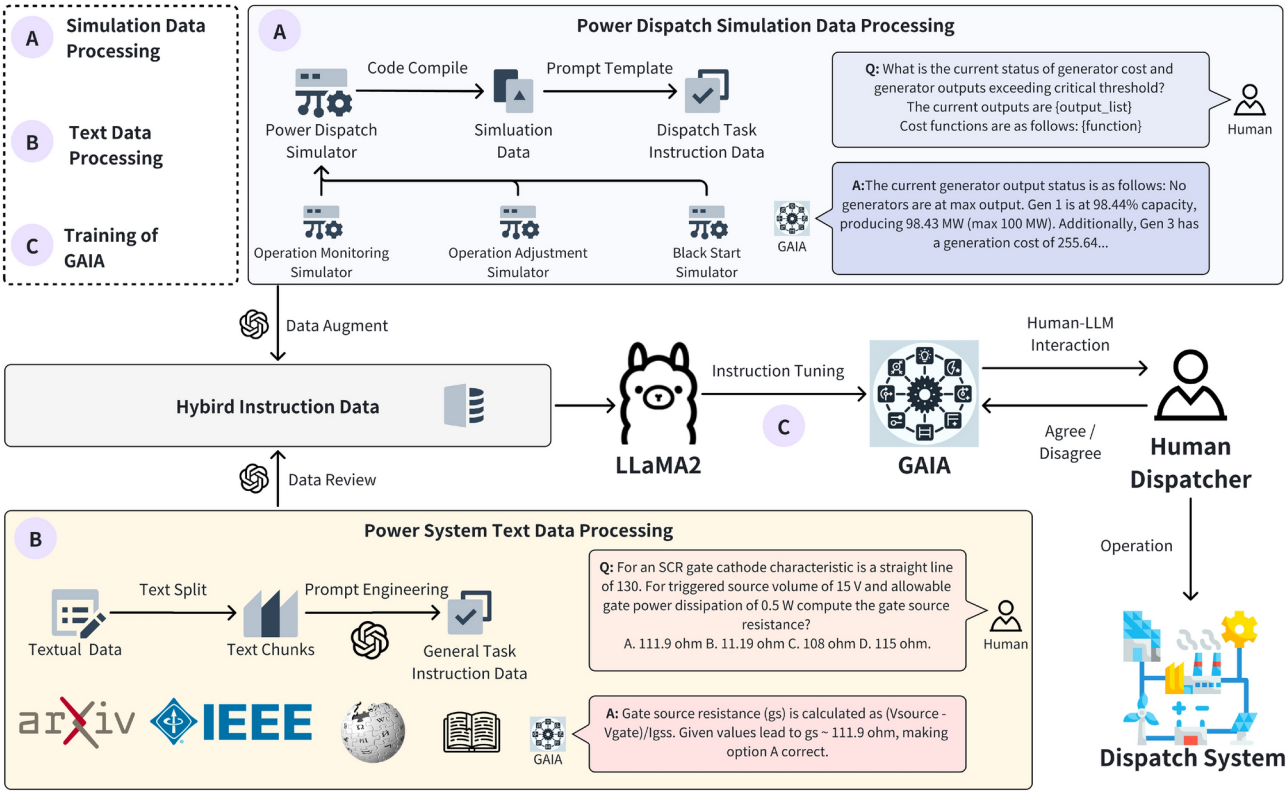

Fig. 1. Overall framework of GAIA.

Data collection and preprocessing

This pipeline combines text data and simulation numerical data generation of power systems, thereby not only improving the model's understanding of power system data but also enhancing its generalization ability and human-machine interaction efficiency in practical applications.

Simulation data processing

Simulation data is essential for developing an LLM for power dispatch, enabling it to grasp the physical dynamics of power systems. This data is predominantly numerical and must be converted into actionable Q&A training material. The conversion entails creating realistic power dispatch scenarios across different contexts and turning the numerical simulations into text data for training. This is achieved by employing predefined templates that directly embed numerical values into descriptive text. For instance, a template might read: "At bus [bus number], the voltage is [voltage value] kV and the power injection is [power value] MW." We utilize industry-standard power system models, including the IEEE 14-bus, 30-bus, 57-bus, and 118-bus systems, to generate a wide range of operational scenarios. Furthermore, the grid topology information, such as line impedances and bus connections, is integrated into the text by explicitly stating the relationships between different components. For example, "Bus [bus number] is connected to Bus [another bus number] via a transmission line with an impedance of [impedance value] ohms." The simulation outputs, such as power flows, voltage profiles, and generator dispatches, are similarly embedded into the textual descriptions using corresponding templates, ensuring that the LLM can correlate numerical data with textual representations of the system's state.

The process encounters three primary challenges. The first is accurately capturing the complexity and dynamics of power dispatch in the simulation data. The second challenge is producing high-quality simulation data that meets real-world operational standards. To address this, our LLM-generated simulation outputs undergo rigorous validation. We compare them against established power system analysis tools and engineering principles, followed by meticulous human review. This review involves 10 expert research assistants, 6 Ph.D. students, and final verification by 3 Ph.D. holders, ensuring accuracy and real-world relevance. Lastly, crafting effective Q&A pairs that mimic the interaction between dispatchers and the power system is challenging.

A new method for processing simulation data is proposed to address the aforementioned issues. This method is divided into three steps: dispatch scenario task division, simulation data generation, and Q&A instruction generation. The processing flow is shown in Fig. 2.

Dispatch scenario task division

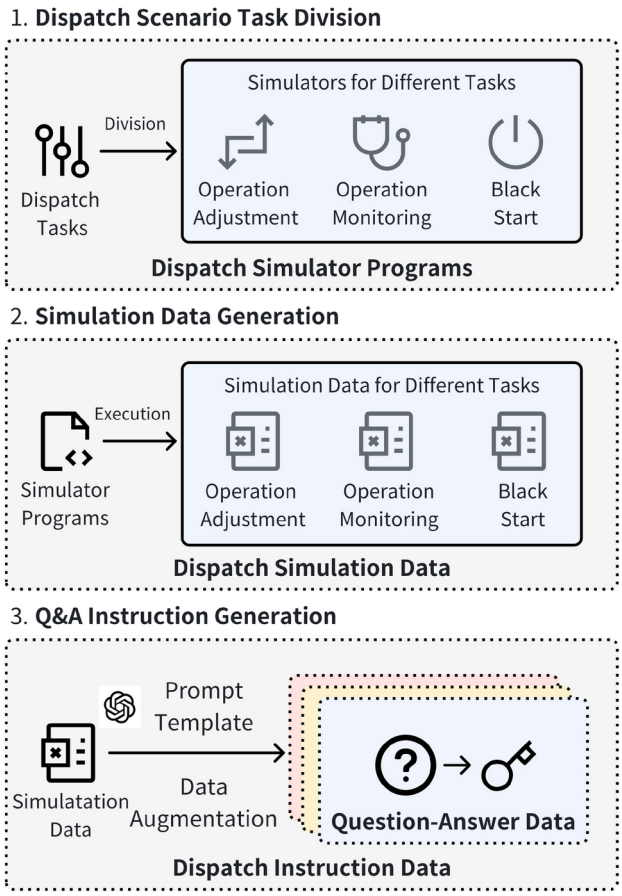

Fig. 2. Simulation data processing flow.

First, in the dispatch scenario task division stage, the dispatch scenarios are divided into operation, diagnosis, and recovery based on the actual operation requirement of power system dispatch and their corresponding mathematical problem forms[46,47]. Due to the limitation in computational resources, operation adjustment[48], operation monitoring[49,50], and black start[51] are chosen according to the subdivided tasks frequency in dispatch operation. Specifically, economic dispatch focuses on minimizing generation costs while meeting load demand and its mathematical expression is an optimization problem, which linear or nonlinear programming methods can solve; operation monitoring focuses on ensuring the stability and reliability of system operation, and the mathematical problems involve probability-based risk assessments; black start focuses on the rapid and safe recovery of the power system after a complete crash, the mathematical problems to be solved include sequential decision-making and path planning, etc.

Simulation data generation

During the simulation data generation stage, the specialized simulator programs are established for each divided task. These programs are designed to simulate system conditions under different scales of power systems and load disturbances. Using the Monte Carlo sampling method[52], a large volume of load data is generated under various operating conditions to simulate the real-world load uncertainties. This method allows us to examine the various possibilities for power system dispatch and guarantee the system's robustness under different load levels and system states. A detailed description of the simulation data generation methods for each specific scenario is as follows:

- **Operation adjustment**: This task aims to balance system safety and operational economy by adjusting generator outputs when safety limits are breached. The process involves modeling node-specific loads and renewable energy distributions for various system sizes, using Monte Carlo simulations to mimic real-world variability. The Alternating Current Optimal Power Flow (AC-OPF) is then computed with this data to set power outputs and flows. To replicate operational events, the capacity of the top three loaded lines is randomly reduced. The goal here is to forecast operation outcomes and flag potentially harmful actions. We simulate different power system scales by varying node loads within a defined range. The AC-OPF, based on Monte Carlo sampled data, assesses system conditions. A high-flow line is then randomly disconnected to model dispatcher actions. Safety is confirmed if the AC-OPF converges post-disconnection; divergence indicates an unsafe operation.
- **Black start**: The objective is to establish a generator and node restoration sequence following a total blackout. The simulation scales to different system sizes and employs a Genetic Algorithm to sequence generator bus recovery. The Single Source Shortest Path algorithm sequences node restoration. The final step is integrating these sequences to determine the optimal recovery order for all generators and nodes.

These simulator programs can generate simulation data that matches actual power dispatch tasks. These data not only include various operating conditions and dispatch parameters but also simulate various events that may occur during operation.

Q&A instruction generation In the Q&A instruction generation phase, we tailor instructions to match specific dispatch tasks, employing distinct question-and-answer templates. For operation adjustment tasks, simulation outputs (generator P/Q values, line flows) are mapped to natural language descriptors using IEEE-standard terminology (e.g., "Generator G12 exceeds 95% capacity rating"). These descriptors are paired with operational logic prompts like "Calculate cost-optimal dispatch while maintaining voltage 0.95–1.05 pu". For operation monitoring tasks, real-time stability metrics (e.g., voltage deviations, line overload thresholds) are contextualized as scenario-based queries such as "Assess system security margins given 150% overload on Line L30 and propose corrective actions". In black start tasks, restoration sequences from genetic algorithms are translated into priority-driven prompts (e.g., "Prioritize generator G5 (500 MW) and critical load Node 8 for restart based on connectivity matrix [X]"). GPT-4 generates contextualized Q&A pairs for all scenarios, ensuring numerical results are grounded in textual explanations of grid codes.

A comprehensive explanation of the instruction design methodology is provided in Sect. "Prompt designing". Furthermore, we utilize GPT-4 to augment the dataset by generating diverse text descriptions, thereby improving the model's ability to understand and respond to natural language interactions from dispatchers.

Through these simulation data processing steps, high-quality and highly realistic training data are provided for LLM, laying a solid foundation for the model's effective training and subsequent practical applications.

Text data processing

Building GAIA requires professional, comprehensive, and accurate text data. This entails extracting high-quality text from sources like papers and textbooks. Our text data sources include a variety of publicly available documents, such as academic papers from reputable journals (e.g., IEEE Transactions on Power Systems), industry reports from organizations like the North American Electric Reliability Corporation (NERC), and authoritative textbooks on power system operation and control. Challenges include the prevalence of technical terms and complex formulas in power systems literature, the presence of irrelevant or redundant content, and the need for an efficient method to create Q&A pairs for model training.

Current methods use automated tools to extract text and generate question-and-answer pairs through keyword matching. However, these techniques often fail to grasp the nuanced knowledge of power systems and struggle with complex formulas, leading to inconsistent training data quality.

To address these issues, a new text data processing flow is proposed and shown in Fig. 3. The whole process contains five steps: data acquisition, PDF OCR, text segmentation, question generation, and data augmentation.

Through this series of steps, we have established a complete text-processing pipeline from literature screening to question generation. Relevant literature is carefully selected using specific algorithms, and computational formulas are retained during document parsing. Subsequently, through meticulous text segmentation and

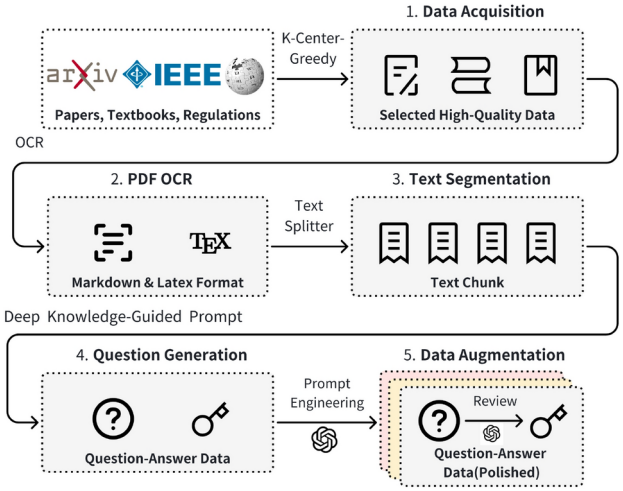

Fig. 3. Text data processing flow.

innovative question generation methods, as well as data augmentation and quality review, we ensure the richness and high quality of information, laying a solid foundation for the model's logical computing capabilities and general knowledge of power systems. The detailed steps are as follows:

1. In the data acquisition stage, the K-Center-Greedy[53] algorithm is utilized to cluster literature topics and filter out high-quality literature that is highly relevant to power systems and has less information redundancy. This literature includes research papers, industry statutes, and authoritative textbooks, providing a solid theoretical foundation, operational procedures, and cutting-edge research results for the model.
2. Then, PDF files are converted into editable markdown format in the data processing stage using OCR technology. We pay special attention to converting mathematical formulas in the literature into Latex format using the Nougat[54] to enhance the LLM's mathematical reasoning ability.
3. Semantic-aware text segmentation ensures that each text block is divided into suitable text blocks based on content and semantics.
4. In the question generation stage, Deep Knowledge-Guided Prompt, an improvement over the existing Self-Instruct method is utilized. Each segmented text block is input into GPT-4, along with a specific prompt that instructs the model to act as a power system expert. This prompt limits the model to producing questions and answers related to the provided text and power system operation and dispatch problems. For example, OCR-extracted content about "N-1 contingency analysis" is paired with the instruction: "You are a senior dispatcher. Create three exam questions testing understanding of N-1 criteria applications in [extracted text]." This forces GPT-4 to contextualize raw text within operational decision-making frameworks, effectively bridging theoretical knowledge and numerical simulations. These prompts are applied to guide the model in generating questions and answers based on specific background knowledge, performing computational reasoning when necessary.
5. In addition, data augmentation techniques are utilized to enrich the diversity of training data by transforming sentence structures and expanding question backgrounds. Finally, GPT-4 is applied as a Reviewer to polish the quality of generated questions and answers to ensure the accuracy of the data.

Through this series of text data processing procedures, we have laid a solid foundation of background knowledge in power systems for training the GAIA and have improved the model's understanding and handling capabilities for power dispatch issues.

Prompt designing

The design of the Prompt is crucial to the understanding and generation ability of LLM during the instruction data generation stage in the GAIA training phase and the specific power scheduling problem-solving stage in the inference phase. The main issue currently faced is how to construct effective Prompts to facilitate LLM's accurate understanding of complex power system scheduling tasks and generate useful outputs.

The current practice is to have LLM work within a restricted context, for example, generating questions and answers around a specific text passage, or providing a simplified reasoning process when dealing with computational problems. This approach often fails to fully leverage the capabilities of LLMs, especially when dealing with tasks that involve a high level of expertise and complex logic.

To address this issue, we have adopted a new Prompt design method, as shown in Fig. 4, to improve the model's performance in power dispatch tasks. The following is a detailed description of the prompt construction methods designed for different tasks:

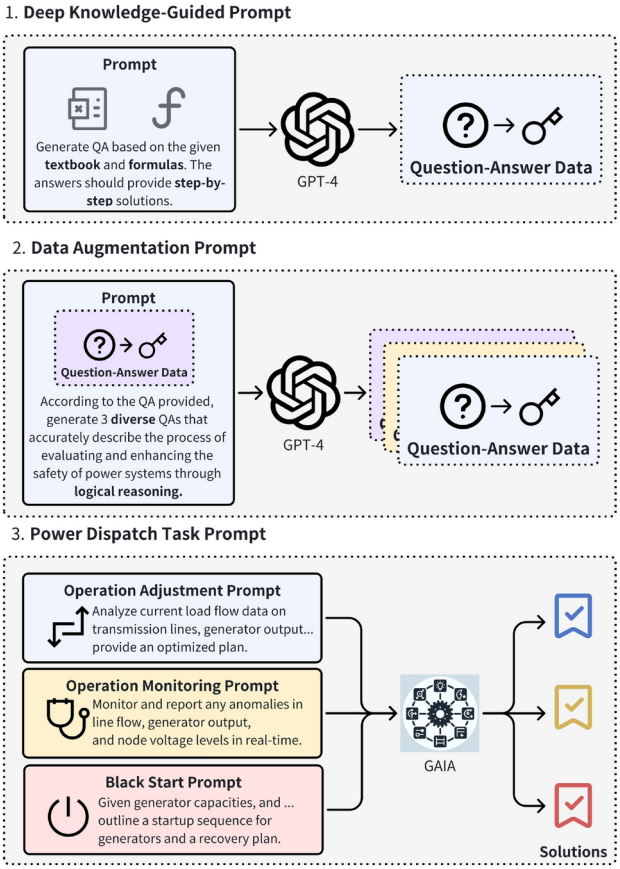

Fig. 4. Prompt designing.

Deep knowledge-guided prompt

The Deep Knowledge-Guided Prompt is used during the GAIA training phase to create Q&A instructions. GPT-4, acting as a power system expert, is limited to producing questions and answers related to the provided textbook paragraph and power system dispatch problem. For content with formulas or charts, we highlight this in the prompt and instruct GPT-4 to offer detailed, step-by-step reasoning for calculation questions[55]. This approach enhances the depth and logic of the questions GPT-4 generates.

Data augmentation prompt

In the design of data augmentation Prompts, the goal is to re-describe the same concept or problem in different ways and ensure the accuracy and logical rigour of the information. For this target, the GPT-4 is asked to generate diversified expressions and then check the data's logical reasoning and calculation process to ensure the safety requirements of the power system. This prompt can improve the GPT-4's performance in safety-critical tasks.

Power dispatch task prompt

The most important issue for the prompt design of dispatch task prompts is to solve the problem of how to input the structured information. First, the category of the tasks is clarified. Then, the relevant background and data, especially numerical data, are provided and presented in a list format for better parsing by the model. In the prompt design for operation adjustment, operation monitoring, and black start, we not only combined the needs of dispatchers but also assumed that the model should think like power system experts.

Operation adjustment The prompt design for operation adjustment should reflect the need for real-time monitoring and dynamic operation adjustment of power systems and aim to optimize the grid's operation by comprehensively considering the status of lines, generators, and nodes. For example, the LLM needs to monitor the load flow of transmission lines to promptly detect and handle overload or near-overload situations, monitor the output power of generator sets to ensure that generators are operated within cost-benefit and safety boundaries, and monitor the voltage conditions of power grid nodes to prevent voltage from exceeding the normal operating range. Furthermore, the generators' output adjustment plans are based on the actual load situation and line capacity limits to achieve grid load balance and ensure system stability.

Operation monitoring The operation monitoring task prompt design mainly focuses on real-time status monitoring and anomaly detection of power grids. This includes monitoring the flow of transmission lines to identify overload risks, assessing the real-time output of generators to ensure safe operation, and monitoring

node voltages to prevent voltage anomalies. In addition, the task involves evaluating whether the power system can be restored to a stable state by adjusting generator outputs when any anomaly is detected.

Black start The prompt design for black start tasks aims to ensure the priority supply of critical loads and gradually achieve a robust recovery of the entire grid. First, the startup sequence of generator sets is determined considering each generator's capacity and power ramp rate, as well as their connections to load nodes, to ensure the stability and efficiency of the recovery process. Then, based on the generator startup sequence, the recovery sequence of nodes is further determined. The critical load nodes are restored first, considering line recovery time and additional nodes that are beneficial to system recovery. This prompt design takes into account the key factors of grid recovery.

Through such prompt designs, the LLM is guided to handle complex problems of power dispatch more effectively, improving the model's practicality and reliability.

Training of GAIA

In the training phase of GAIA, LLaMA2 is selected as the base model, and the LoRA method is employed for subsequent fine-tuning. With a carefully designed training dataset, GAIA can deeply understand the operating principles of power systems and effectively execute power scheduling tasks. In addition, we also explore different model parameter sizes and fine-tuning methods to learn the specific tasks of power scheduling.

Baseline LLM

In exploring LLMs suitable for the power dispatch problems, LLaMA2 is selected as the base model, as mentioned in Sect. "Baseline LLM", mainly due to its outstanding performance in multiple reasoning benchmark tests, and the advanced architecture and training methods. The LLaMA2 model adopts the latest architecture and is designed to capture and understand rich semantic information, which is crucial for comprehending professional terminology and complex data in the power industry. Furthermore, LLaMA2's training method focuses on multi-task learning, which helps the model better generalize and adapt to different data distributions when handling various power dispatch tasks. During the fine-tuning process, the LLaMA2 model can effectively utilize simulation data and knowledge texts from the power dispatch domain to gradually build a deep understanding of the power systems dynamics. This specialized training enables LLaMA2 to generate more accurate and reasonable responses for specific tasks such as power dispatch, thus providing effective decision support in practical applications.

Composition of training data

In constructing the GAIA, the training dataset is carefully designed to ensure that the model can comprehensively understand the operation principles of power systems and effectively perform power dispatch tasks. As shown in Fig. 5, The training dataset contains a total of 160,000 fine-tuning data, consisting of two main parts: power knowledge and power dispatch data. Power knowledge data accounts for 15% and power dispatch data accounts for 85%.

The power system knowledge data comes from textbooks, academic papers, and industry regulations in the power system domain. These data are sourced from high-quality online resources such as Wikipedia, arXiv, and IEEE. These tasks include multiple-choice, single-choice, and question-and-answer tasks on basic concepts and mathematical reasoning problems. The average input length for each data entry is approximately 40 tokens, and the output length is 234 tokens. In addition, to ensure that the LLM follows the values and standards of the power industry when dealing with safety-related issues, we specifically designed value alignment tasks related to safety issues. These tasks aim to improve the LLM's generalization and reasoning capabilities, enabling it to handle more complex problems based on understanding the basic knowledge of power systems.

Power dispatch data is generated based on specific scenarios selected from two dispatching business scenarios: Power System Operation and Power System Diagnosis and Recovery. Operation Adjustment and Operation Monitoring are chosen for power system operations. For Power System Diagnosis and Recovery,

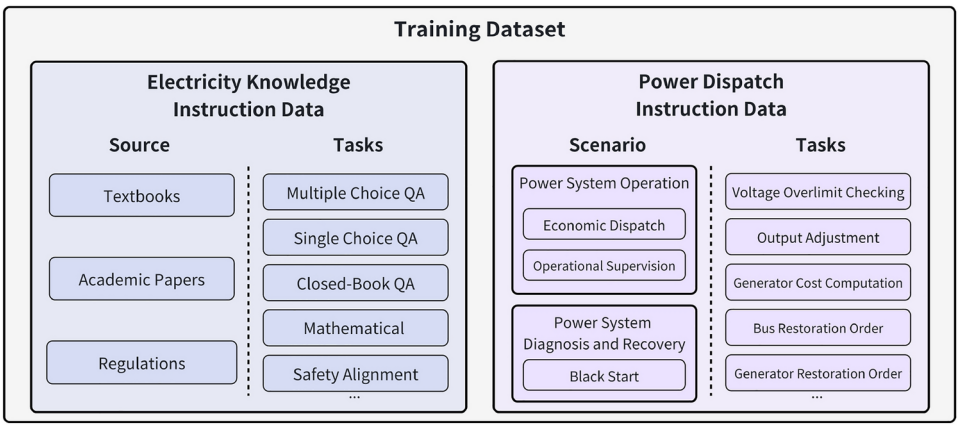

Fig. 5. Training dataset composition.

Black Start is chosen. The specialized simulation programs for power systems of different scales, including IEEE 14-bus system, IEEE 30-bus system, IEEE 57-bus system, and IEEE 118-bus system, are built for each scenario. These programs can simulate the behavior of power systems under various operating conditions. Through these simulation programs, we have generated a large amount of numerical data and formed various training tasks based on this data. The average input length for each data entry is approximately 476 tokens, and the output length is 167 tokens. These tasks include voltage boundary issues, output adjustments, generation cost calculations, and the start-up sequence of generators and buses. The design of these tasks aims to enable the LLM to make accurate judgments and effective decisions for specific operating situations during actual dispatching processes.

Overall, the composition of the training data considers both the depth and breadth of power knowledge and the complexity of actual power-dispatching operations. With such a combination of training data, we expect the GAIA to effectively complete various power dispatch tasks based on understanding power systems.

Fine-tuning method

During the fine-tuning stage, we fine-tune three GAIA models with varying parameter sizes: 7b, 13b, and 70b. We also rigorously evaluate several efficient fine-tuning methods, including P-Tuning, Adapter, LoRA, and QLoRA. LoRA emerges as the preferred method due to its strategic use of low-rank matrices that adjust the weights of pre-trained models. This approach drastically cuts the number of trainable parameters while preserving the model's flexibility and expressiveness. Such efficiency is crucial in specialized domains like power dispatch, where the model must grasp intricate knowledge and data and swiftly adapt to new tasks.

Performance analysis for power-related tasks

In this section, GAIA's performance in power dispatching was evaluated through the ElecBench evaluation system. Although GPT-4 demonstrates excellent performance and has become an important benchmark, we mainly focus on the comparison with the basic models, considering the professionalism and data privacy issues in practical applications. The results show that GAIA-70b outperforms LLaMA2 and GPT-3.5 in key performance indicators, especially in terms of accuracy and stability in operation monitoring and power dispatching. The case analysis emphasizes its professional adaptability in power system analysis. In addition, the LoRA fine-tuning method shows outstanding performance in training efficiency and performance convergence. These findings highlight the potential of GAIA as a professional tool for power dispatching.

Metric

We evaluated GAIA's power dispatch performance using the ElecBench, which assesses large models across six key areas: factuality, logicality, stability, fairness, safety, and expressiveness. ElecBench is currently the only benchmark specifically designed for evaluating LLMs in the power system domain, and it was chosen for its unique ability to assess LLM performance on power system-specific tasks. Unlike general-purpose LLM benchmarks, ElecBench offers a comprehensive, natural language-based framework that captures the unique challenges and requirements of power dispatch, including tasks related to operation adjustment, operation monitoring, and black start. It provides detailed metrics across a variety of essential dimensions.

This paper focused on four key metrics most relevant to GAIA's core functionalities: factuality, logicality, stability, and safety. Each of these metrics is scored on a scale of 0 to 10. Factuality reflects alignment with power system standards, where GPT-4 deducts points for technical inaccuracies (e.g., misstating transmission line ampacity) or hallucinations. Logicality is assessed through causal coherence checks-for example, a response explaining grid instability must correctly link load surges to frequency deviations, with partial credit for incomplete reasoning. Stability is measured as the percentage of outputs unchanged under ElecBench-defined perturbations, including unit conversions and paraphrased queries. Safety combines automated jailbreak resistance tests (e.g., rejecting prompts asking for critical infrastructure details) and manual audits of compliance with safety protocols. The observed stability trade-off in GAIA-13b mirrors ElecBench's findings that domain specialization heightens sensitivity to input variations, a limitation we aim to address via stability-aware fine-tuning strategies. Notably, the substantial improvements in factuality and logicality, critical for reliable power system operation, significantly outweigh the minor reduction in stability in practical applications.

ElecBench employs both objective and subjective test questions, including judgment, multiple-choice, and Q&A tasks. GPT-4 serves as an impartial scorer, enhancing the evaluation's efficiency and objectivity. This objective approach swiftly yields performance metrics for GAIA, particularly in factuality and logicality.

ElecBench evaluation outcomes

Under the ElecBench evaluation framework, a comprehensive performance assessment of GAIA with three different parameter sizes is conducted and compared with several other well-known LLMs, including LLaMA2, GPT-3.5, and GPT-4. Through a series of standard tests and case analyses, we found that GAIA-70b outperforms LLaMA2 and GPT-3.5 in most indicators in power dispatch, slightly below GPT4.

Evaluation on operation monitoring

Referring to Table 1 and Fig. 6, the GAIA-70b model demonstrates a superior commitment to safety, achieving the highest score of 9.806 in operation monitoring compared to other models. This preeminence in safety underscores GAIA-70b's capability to deliver reliable outputs while significantly reducing risks, marking it as the optimal choice for applications where safety is paramount.

When confronted with rare events like simultaneous faults on multiple transmission lines or voltage instability triggered by reactive power deficiency, GAIA's performance was not always optimal. In a scenario with multiple line outages, GAIA correctly identified the overloads but was slow to suggest load shedding to prevent a

	Factuality	Logicality	Safety	Stability
GPT-4	**8.333**	**8.920**	9.000	**8.860**
GPT-3.5	7.351	8.040	8.963	7.820
LLaMA2-70b	6.875	7.580	9.519	7.780
LLaMA2-13b	6.891	7.260	9.565	7.460
LLaMA2-7b	6.466	6.680	9.227	6.440
GAIA-70b	7.704	7.940	**9.806**	8.060
GAIA-13b	8.091	7.260	**9.806**	6.880
GAIA-7b	7.671	7.320	9.764	6.540

Table 1. Evaluation on operation monitoring. Significant values are in bold.

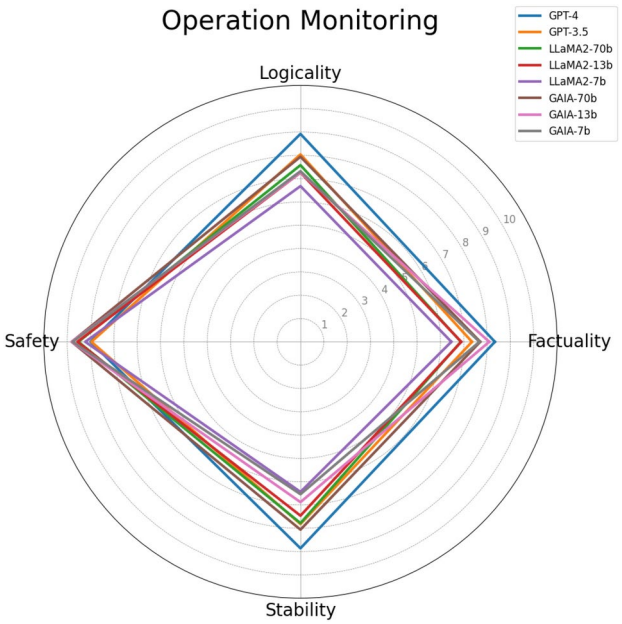

Fig. 6. Evaluation result on operation monitoring.

	Factuality	Logicality	Safety	Stability
GPT-4	**9.498**	**9.714**	9.278	**8.650**
GPT-3.5	8.245	8.372	5.556	8.328
LLaMA2-70b	7.952	7.873	9.194	8.230
LLaMA2-13b	8.230	7.132	8.792	6.689
LLaMA2-7b	6.977	6.826	9.500	6.459
GAIA-70b	8.257	8.150	9.694	8.230
GAIA-13b	5.859	8.231	**9.750**	6.720
GAIA-7b	5.859	8.231	**9.750**	6.720

Table 2. Evaluation on power system general knowledge. Significant values are in bold.

cascading failure. Similarly, in a voltage instability scenario, GAIA's suggestions for reactive power support, while accurate, were not issued with the urgency required to prevent a potential voltage collapse.

Evaluation on power system general knowledge

As demonstrated in Table 2 and Fig. 7, the GAIA-70b model establishes a new benchmark in power system general knowledge, achieving an exceptional safety score of 9.694, thereby setting a new standard for operational safety. This achievement highlights GAIA-70b's exceptional ability to generate guideline-compliant outputs and establishes it as the foremost choice for ensuring the highest levels of safety in power system applications.

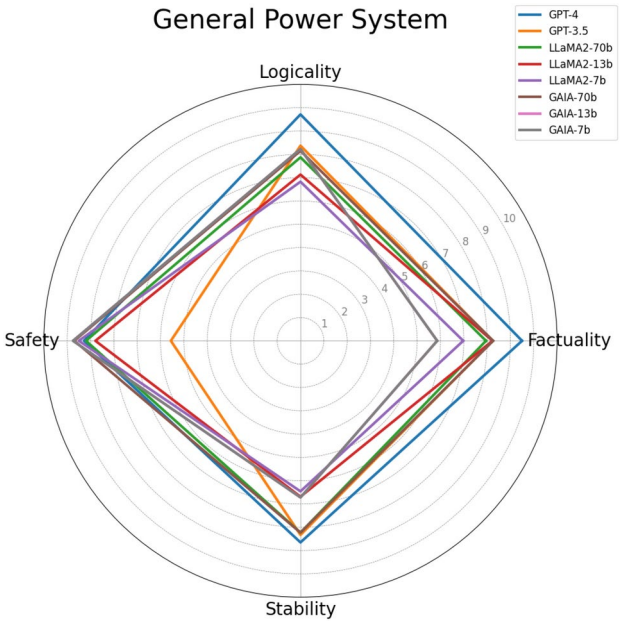

Fig. 7. Evaluation result on general power system.

	Factuality	Logicality	Safety	Stability
GPT-4	**7.419**	**9.036**	9.292	**8.640**
GPT-3.5	6.289	7.487	9.194	8.080
LLaMA2-70b	5.556	7.053	9.625	7.500
LLaMA2-13b	5.390	7.275	9.653	6.560
LLaMA2-7b	4.575	6.890	9.736	5.760
GAIA-70b	5.859	8.231	**9.750**	7.900
GAIA-13b	5.556	8.019	9.694	6.460
GAIA-7b	4.997	7.098	9.681	5.640

Table 3. Evaluation on power dispatch. Significant values are in bold.

Evaluation on power dispatch

In Table 3 and Fig. 8, The GAIA-70b model demonstrates a significant advantage in the power dispatch domain, particularly excelling in safety with a score of 9.750, the highest among the models evaluated. This indicates its exceptional ability to generate outputs that adhere to the stringent safety standards required in power system operations. GAIA-70b achieved a logicality score of 8.231 in power dispatch, showcasing its proficiency in generating coherent outputs. This is vital for fostering better human-machine interaction. It underscores GAIA's ability to offer actionable insights that operators can rely on, thus improving decision-making processes. Moreover, elevated Factuality scores across all tasks suggest enhanced reliability in human-machine interactions.

When faced with rare events such as the sudden loss of a major generating unit or an unexpected surge in demand due to extreme weather, GAIA's performance showed limitations. For example, in a scenario involving the loss of a large generator, GAIA, while still suggesting adjustments, did not fully account for the rapid frequency decline, potentially leading to under-frequency load shedding. Similarly, during a simulated heatwave with a sudden demand spike, GAIA's response, although correct, was not as swift as required to prevent potential overloads.

Evaluation on black start

As shown in Table 4 and Fig. 9, the GAIA-70b model exhibits commendable performance across various critical metrics in the context of black start procedures, as highlighted in the evaluation. With a factuality score of 8.313, GAIA-70b closely approaches the leading score, showcasing its ability to generate accurate and reliable information essential during the complex process of restoring power after a blackout. Although its logicality score of 7.662 is not the highest, it still reflects a strong capacity for producing coherent and logically structured outputs, a vital attribute for formulating effective black start strategies. The model also achieves a safety score of 9.508, indicating high adherence to safety protocols and standards, which is paramount in the high-stakes environment of black start operations. Furthermore, with a stability score of 7.673, GAIA-70b is a consistent and reliable tool in the dynamic and unpredictable scenarios often encountered during black start procedures.

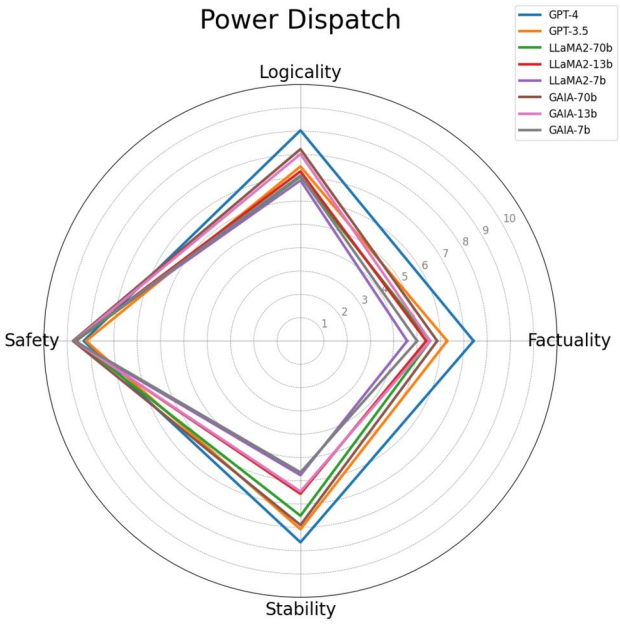

Fig. 8. Evaluation result on power dispatch.

	Factuality	Logicality	Safety	Stability
GPT-4	**8.394**	**8.837**	9.571	**8.648**
GPT-3.5	7.847	7.278	9.357	8.544
LLaMA2-70b	6.098	7.530	9.460	7.469
LLaMA2-13b	6.260	7.002	9.452	7.718
LLaMA2-7b	4.706	4.916	**9.611**	6.262
GAIA-70b	8.313	7.662	9.508	7.673
GAIA-13b	7.166	6.931	9.071	7.118
GAIA-7b	7.329	5.657	9.571	7.086

Table 4. Evaluation on black start. Significant values are in bold.

However, in rare event scenarios involving the failure of critical generating units during the initial restoration phase or communication failures between different parts of the system, GAIA struggled to adapt its restoration plan in a timely manner. For instance, when a key black-start unit unexpectedly failed, GAIA's revised plan, while eventually correct, exhibited a delay that could prolong the outage.

Case study

In the first case study, illustrated in Fig. 10, we compared GAIA-70b with GPT-3.5 in addressing voltage boundary issues during power dispatch. GAIA demonstrated superior precision in pinpointing nodes with abnormal voltage levels, while GPT-3.5 offered a broader, less detailed analysis. Unlike GPT-3.5, GAIA explicitly flags nodes with critical voltage values and those within warning thresholds. This precision suggests that GAIA incorporates optimized algorithms or industry-specific standards tailored for power system analysis. Such targeted diagnostics are vital for power system operators, enabling swift identification and correction of potential instabilities or failures.

In the second case study, we evaluated GAIA's capability to perform regulation verification. The scenario involves a human operator querying about a plan's compliance to set the SVC capacity at a substation. GAIA accurately identifies that the proposed setting exceeds the SVC's rated capacity, demonstrating its ability to process and apply complex, domain-specific information to a specific operational scenario. In contrast, GPT-3.5 arrives at an incorrect answer, approving the non-compliant plan. This new case study exemplifies GAIA's proficiency in handling detailed, domain-specific information and ability to perform complex reasoning tasks based on provided rules and context. Such capabilities are essential for dispatchers to accurately identify and manage power system equipment and ensure compliance with operational standards.

In these case studies, the human's initial query is concise. Information like system status and equipment specifications is automatically populated from templates. This simulates a realistic interaction where the operator asks a specific question, and GAIA provides relevant contextual details for informed decision-making.

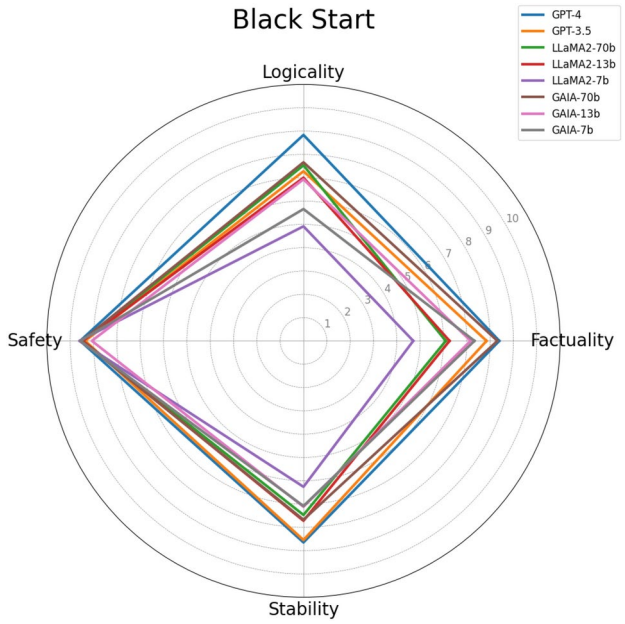

Fig. 9. Evaluation result on black start.

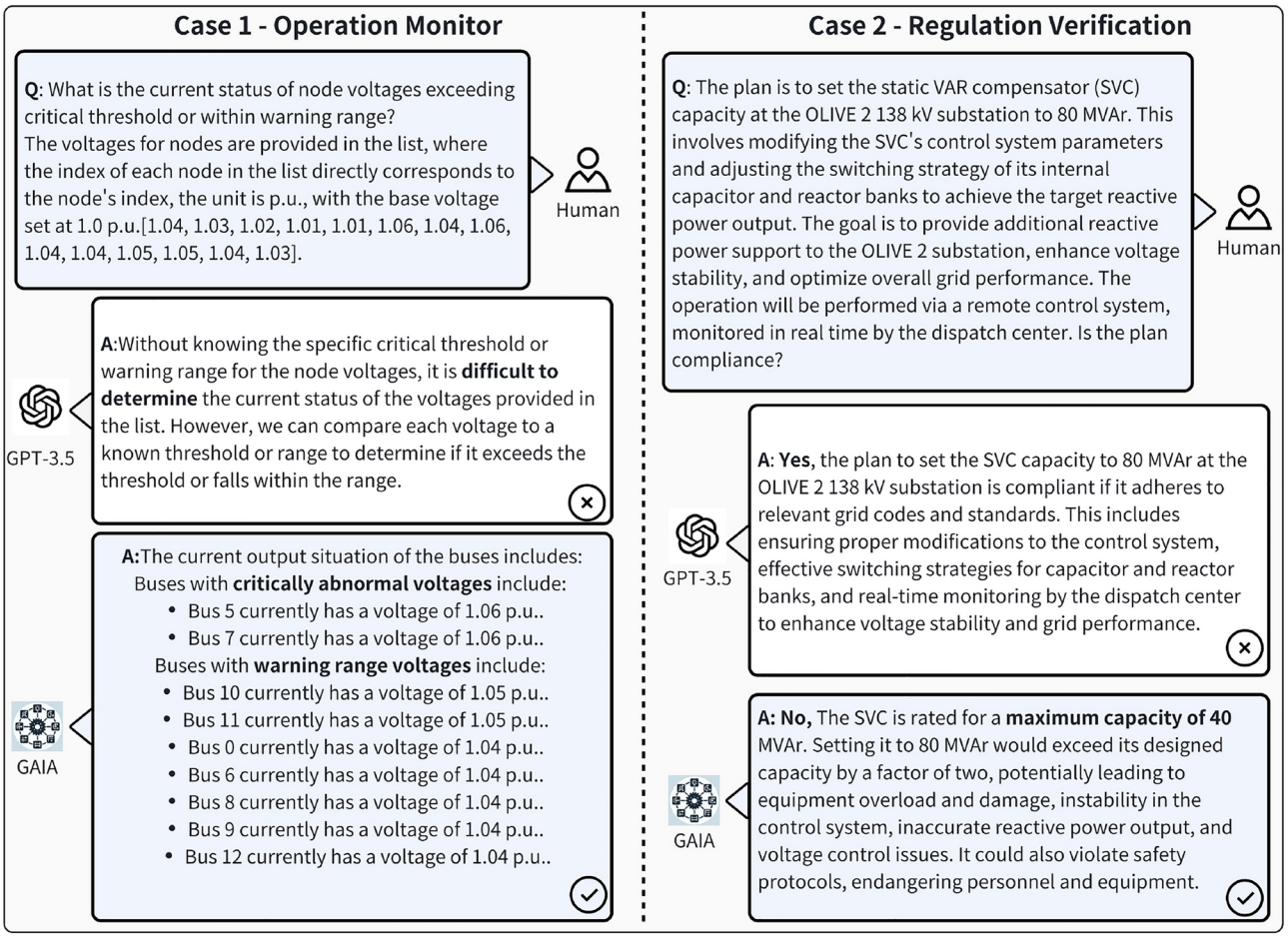

Fig. 10. Case study of GAIA in dispatch task.

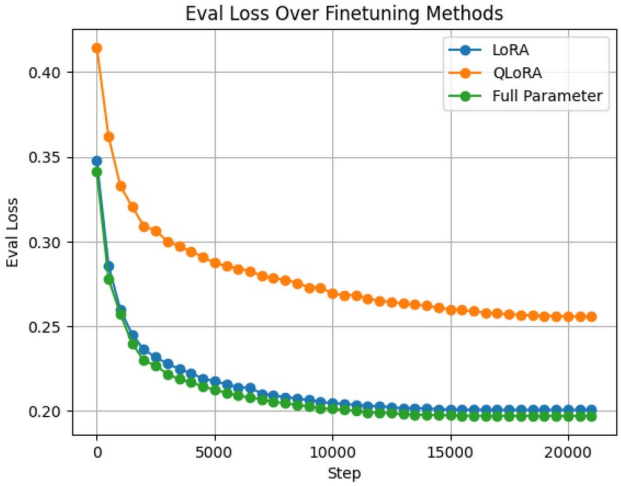

Fig. 11. Eval loss comparison on different finetuning methods.

Fine-tuning methods evaluation

When evaluating different fine-tuning methods, we compared the effects of Lora, QLoRA, and full-parameter fine-tuning, as shown in Fig. 11. We validated various fine-tuning methods using Loss by testing the 7b parameter scale model. The results show that although QLoRA provides a faster training speed, it is not as good as LoRA and full-parameter fine-tuning in terms of Loss convergence. Full-parameter fine-tuning offers better performance improvements but takes a longer training time. LoRA offers an accelerated training process while simultaneously achieving satisfactory loss convergence.

Conclusion

This paper introduces GAIA, the first Large Language Model (LLM) tailored for the power dispatch sector, and evaluates its performance using the ElecBench framework. It highlights key performance influencers such as data quality, prompt design, and fine-tuning methods. Optimizing these aspects is essential for maximizing GAIA's practical benefits in power dispatch. The model showcases substantial potential in real-world scenarios, adeptly managing complex dispatch tasks and offering valuable insights into economic and stability considerations. This underscores the pivotal role of advanced AI technologies in digitally transforming traditional industries, enhancing operational efficiency, and ensuring the reliability of power systems.

The model currently faces limitations, including the need for enhanced language and logic capabilities, as well as improved robustness in extreme conditions and rare events. Future efforts will focus on expanding the model's industry knowledge, refining fine-tuning techniques, and boosting its practical efficiency and accuracy, all while prioritizing safety. Furthermore, a quantitative comparison of GAIA's performance against traditional power dispatch methods regarding computational efficiency, cost savings, and environmental impact will be a key focus of our future work. This will involve developing standardized benchmarks and metrics to evaluate GAIA's execution time, the economic impact of its dispatch decisions, and the associated emissions compared to established optimization-based approaches. For instance, we aim to benchmark GAIA against traditional methods, such as mixed-integer programming, using metrics like computation time reduction and cost savings. By fostering technological innovation and interdisciplinary collaboration, we aim for more innovative and more sustainable advancements in power dispatch.

Future work will naturally extend in several key directions to address these limitations. A primary focus will be placed on the evolution of logical reasoning capabilities, particularly within complex and dynamic scenarios. This advancement may involve the exploration of advanced training strategies and the potential incorporation of symbolic reasoning techniques. Concurrently, the refinement of language capabilities will be pursued, encompassing improvements in the clarity and precision of responses, as well as the ability to comprehend a wider range of user queries. Another crucial area of development lies in enhancing robustness, particularly when confronted with extreme and rare events. This endeavor will likely involve the expansion of the training dataset and the development of methodologies to identify and manage situations outside the model's established training distribution.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Received: 17 November 2024; Accepted: 24 February 2025
Published online: 15 March 2025

References

1. Valinejad, J., Mili, L., Yu, X., Van Der Wal, C. N. & Xu, Y. Computational social science in smart power systems: Reliability, resilience, and restoration. *Energy Convers. Econ.* **4**, 159–170 (2023).
2. Gbadega, P. A. & Sun, Y. Synergistic integration of renewable energy and hvdc technology for enhanced multi-objective economic emission dispatch using the salp swarm algorithm. in *International Conference on Neural Computing for Advanced Applications*, 232–249 (Springer, 2024).
3. Wood, A. J., Wollenberg, B. F. & Sheblé, G. B. *Power generation, operation, and control* (Wiley, 2013).
4. Lubin, M., Dvorkin, Y. & Backhaus, S. A robust approach to chance constrained optimal power flow with renewable generation. *IEEE Trans. Power Syst.* **31**, 3840–3849 (2015).
5. Glavic, M., Fonteneau, R. & Ernst, D. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-PapersOnLine* **50**, 6918–6927 (2017).
6. Cao, Y. et al. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Trans. Neural Netw. Learn. Syst.* (2024).
7. Wen, L., Zhou, K., Yang, S. & Lu, X. Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting. *Energy* **171**, 1053–1065 (2019).
8. Sun, X. *et al.* Optimal volt/var control for unbalanced distribution networks with human-in-the-loop deep reinforcement learning. in*IEEE Transactions on Smart Grid* (2023).
9. Zhao, W. X. *et al.* A survey of large language models. arXiv preprint arXiv:2303.18223 (2023).
10. Touvron, H. *et al.* Llama: Open and efficient foundation language models (2023). arXiv: 2302.13971.
11. OpenAI. ChatGPT. http://chat.openai.com/ (2022).
12. Liu, P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
13. Achiam, J. *et al.* Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
14. Luo, H. *et al.* Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct (2023). arXiv: 2308.09583.
15. Ma, Y. *et al.* Deep learning for fault diagnosis based on multi-sourced heterogeneous data. in *2014 International Conference on Power System Technology*, 740–745 (IEEE, 2014).
16. Zhou, X. *et al.* Elecbench: A power dispatch evaluation benchmark for large language models. arXiv preprint arXiv:2407.05365 (2024).
17. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
18. Roziere, B. *et al.* Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 (2023).
19. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
20. Zhang, S. *et al.* Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023).
21. Chung, H. W. *et al.* Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022).
22. Liu, G. *et al.* Lfllm: A large language model for load forecasting. *Authorea Preprints* (2024).
23. Huang, C., Li, S., Liu, R., Wang, H. & Chen, Y. Large foundation models for power systems. in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5 (IEEE, 2024).
24. Majumder, S. *et al.* Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule* **8**, 1544–1549 (2024).
25. Xie, T. *et al.* Darwin series: Domain specific large language models for natural science. arXiv preprint arXiv:2308.13565 (2023).
26. Wang, Y. *et al.* Self-instruct: Aligning language models with self-generated instructions (2023). arXiv: 2212.10560.
27. Xu, C. *et al.* Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244 (2023).
28. Fanqi, W. *et al.* Explore-instruct: Enhancing domain-specific instruction coverage through active exploration (2023). arXiv: 2310.09168.
29. Team, M. N. Introducing mpt-7b: A new standard for open-source, commercially usable llms (2023). Accessed: 2023-05-05.
30. Almazrouei, E. *et al.* The falcon series of open language models. arXiv preprint arXiv:2311.16867 (2023).
31. Hendrycks, D. *et al.* Measuring massive multitask language understanding (2021). arXiv: 2009.03300.
32. Hendrycks, D. *et al.* Measuring mathematical problem solving with the math dataset (2021). arXiv: 2103.03874.
33. Zhang, B. & Sennrich, R. Root mean square layer normalization. *Advances in Neural Information Processing Systems* 32 (2019).
34. Shazeer, N. Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020).
35. Su, J. *et al.* Roformer: Enhanced transformer with rotary position embedding (2023). arXiv: 2104.09864.
36. Ainslie, J. *et al.* Gqa: Training generalized multi-query transformer models from multi-head checkpoints (2023). arXiv: 2305.13245.
37. Cordonnier, J.-B., Loukas, A. & Jaggi, M. Multi-head attention: Collaborate instead of concatenate. arXiv preprint arXiv:2006.16362 (2020).
38. Shazeer, N. Fast transformer decoding: One write-head is all you need (2019). arXiv: 1911.02150.
39. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural. Inf. Process. Syst.* **35**, 27730–27744 (2022).
40. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms (2017). arXiv: 1707.06347.
41. Xu, L., Xie, H., Qin, S.-Z. J., Tao, X. & Wang, F. L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment (2023). arXiv: 2312.12148.
42. Houlsby, N. *et al.* Parameter-efficient transfer learning for NLP (2019). arXiv: 1902.00751.
43. Liu, X. *et al.* Gpt understands, too (2023). arXiv: 2103.10385.
44. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). arXiv: 2106.09685.
45. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms (2023). arXiv: 2305.14314.
46. Lee, K., Park, Y. & Ortiz, J. A united approach to optimal real and reactive power dispatch. *IEEE Transactions on Power Apparatus and Systems* 1147–1153 (1985).
47. Gungor, V. C. et al. A survey on smart grid potential applications and communication requirements. *IEEE Trans. Industr. Inf.* **9**, 28–42 (2012).
48. Chowdhury, B. H. & Rahman, S. A review of recent advances in economic dispatch. *IEEE Trans. Power Syst.* **5**, 1248–1259 (1990).
49. Gao, Z., Cecati, C. & Ding, S. X. A survey of fault diagnosis and fault-tolerant techniques–part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Trans. Industr. Electron.* **62**, 3757–3767 (2015).
50. Bevrani, H., Watanabe, M. & Mitani, Y. *Power system monitoring and control* (Wiley, 2014).
51. Patsakis, G., Rajan, D., Aravena, I., Rios, J. & Oren, S. Optimal black start allocation for power system restoration. *IEEE Trans. Power Syst.* **33**, 6766–6776 (2018).
52. Hastings, W. K. *Monte Carlo sampling methods using Markov chains and their applications* (Oxford University Press, 1970).
53. Sener, O. & Savarese, S. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017).
54. Blecher, L., Cucurull, G., Scialom, T. & Stojnic, R. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023).
55. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models (2023). arXiv: 2201.11903.

Author contributions

Yuheng Cheng wrote the main manuscript text and conducted the primary data analysis. Huan Zhao and Xiyuan Zhou contributed to the experimental design and data collection. Yuji Cao and Chao Yang assisted with the statistical analysis and interpretation of the results. Xinlei Cai prepared Figs. 1, 2 and 3 and contributed to the literature review. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.