Expert Systems With Applications 296 (2026) 128917

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Large-model-based smart agent for time series anomaly detection in power systems

Bingrui Wang[a], Yuan Zhou[a,*], Leijiao Ge[a], Sun-Yuan Kung[b]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China
[b] Department of Electrical Engineering, Princeton University, Princeton, NJ, 08540, USA

ARTICLE INFO

Keywords:
Large model
Smart agent
Anomaly detection
Time series
False data injection
Power system

ABSTRACT

Anomaly detection in time series is critical to ensure the safe and stable operation in power systems. Existing methods face dual challenges of data scarcity and limited interpretability. To address this, we propose a novel large-model-based series-detection development framework named SLEP, leveraging the powerful pre-trained knowledge transfer capabilities of large models to mitigate data scarcity and their natural language understanding/generation abilities to enhance the interpretability of anomalies. Additionally, we introduced a brand-new prompt design template called BRIDOR introduced to further control the interaction quality with the large model. This work represents the first attempt to apply large models to anomaly detection for electric time series data. Extensive experiments on the high-dimensional, multivariate, and data-sparse real-world cases, demonstrate the superior performance of our proposed method. This validates the framework's robustness and pioneers a new paradigm for trustworthy artificial intelligence technology in related fields.

1. Introduction

The rapid development of smart grids and the large-scale integration of emerging renewable energy sources such as wind and solar power have led to a significant increase in the complexity of the structure and operational modes of the power system (Kataray et al., 2023). Against this background, the power system will generate vast amounts of time series data during its daily operation. These electric time series data (ETSD) (Bhattarai et al., 2019) reflect detailed records of the changes in various key parameters in various links of power generation, transmission, and distribution, reflecting not only the real-time operational status of the power system but also revealing the inherent characteristics and fundamental laws of grid operation. Therefore, by monitoring and analyzing these ETSD in real-time (Dong et al., 2023; Li et al., 2023), it is possible to rapidly identify various potential issues in the power system, such as performance degradation or failure of power equipment, localized power outages, and illegal power theft by users. The early detection of these anomalies provides valuable time windows for power departments to take necessary maintenance and intervention measures, thus preventing the expansion and spread of power accidents and ensuring the stable operation and power supply security of the entire power system.

Various methods and techniques have been widely used and studied in the field of ETSD anomaly detection (Musleh et al. 2020, Cooper et al. 2023, Susto et al. 2018, Zidi et al. 2023), which can be mainly divided into traditional hand-crafted methods and learning-based methods. Traditional methods (Alencar et al., 2022; Ashok et al., 2020; Ho et al., 2020; Liu et al., 2021) rely heavily on manual design for specific scenarios, making them inflexible and poorly adaptable to the complexity of modern power grids. Consequently, their applicability is diminishing. Learning-based methods (Ahmed et al., 2019; Alshehri et al., 2024; Ceci et al., 2020; Gholami et al., 2024; Habbak et al., 2023; Jindal et al., 2016; Shabad et al., 2021; Takiddin et al., 2023, 2022; Zanetti et al., 2019; Zhang et al., 2022), on the other hand, are data-driven, can achieve end-to-end automatic modeling and reduce subjective intervention, greatly promoting the development of ETSD anomaly detection problems. However, they face two critical limitations. First, due to the sensitivity of the electricity industry and the associated data privacy concerns, it is often challenging to obtain high-quality, publicly available power datasets that are suitable for training learning algorithms. This results in a data scarcity dilemma that limits the development of learning-based methods. Second, a severe lack of interpretability plagues learning methods represented by unsupervised and deep methods, which tend to operate like "black boxes" and fail to provide clear,human-understandable explanations for why particular instances are flagged as anomalous, thereby undermining trust and diagnostic utility.

To address these dual challenges of data scarcity and interpretability, the use of pre-trained large models (Hadi et al., 2023; Min et al., 2023; Zhao et al., 2023) emerges as a highly promising direction. Large models

* Corresponding author.
E-mail addresses: wangbingrui@tju.edu.cn (B. Wang), zhouyuan@tju.edu.cn (Y. Zhou), legendglj99@tju.edu.cn (L. Ge), kung@princeton.edu (S.Y. Kung).

https://doi.org/10.1016/j.eswa.2025.128917
Received 23 April 2025; Received in revised form 30 June 2025; Accepted 3 July 2025
Available online 9 July 2025
0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

exhibit incredibly complex network structures, typically involving billions or even more parameters. Leveraging such immense scale and being trained on massive and diverse datasets (Radford et al., 2019), large models demonstrate robust representational learning and generalization capabilities, possessing the ability to understand intricate patterns and subtle differences embedded in data from different scenarios (Brown et al., 2020), enabling effective knowledge transfer to downstream ETSD anomaly detection tasks and mitigating the impact of domain-specific data scarcity. Furthermore, large models advanced natural language understanding and generation abilities provide a unique pathway to deliver intuitive, human-readable explanations for detected anomalies, significantly promoting actionable diagnostics and trust in critical power system operations.

The large model is essentially an information processing unit that lacks the necessary abilities of memory, perception, and action, and cannot be directly applied to the complex grid environment. It is necessary to construct a smart agent (Ruan et al., 2023; Wang et al., 2024; Xi et al., 2025), which is an intelligent entity (software and/or hardware) capable of sensing the environment, making decisions, and executing actions, in order to introduce the large model into the power system and accomplish specific tasks. Currently, the most popular agent framework for large models is CVP (ChatGPT-VectorDB-Prompt) (Luan, 2023). It provides an efficient development path of commercial applications for large models by combining the analytical capabilities of ChatGPT, the storage capabilities of vector databases, and specific business-oriented prompts. However, the ETSD detection problem usually requires specific algorithms to pre-process the series generated by the equipment, and timely alerting or processing when anomalies are detected. This is beyond the reach of existing agent frameworks. Therefore, we propose a large-model-based agent framework named SLEP, which is tailored to the series detection problem and can be quickly migrated to any grid scenario that generates time series data.

As illustrated in Fig. 1, S represents the time Series generated in real-time by the power equipment. L refers to the pre-trained Large models that serve as the brain of the framework for anomaly detection. E is the Enhancing module, aiming to support prompts via historical information or the large model's reflection, enabling large models to generate more responsive results. P stands for Prompt, which is the guiding text provided by users to interact with large models. It is the interaction module in the SLEP framework, linking the S and L modules. Besides, we also propose a new prompt design template called BRIDOR to further control the quality of prompts. Extensive experiments conducted on three publicly available real-world ETSD cases, namely "Individual Household Electric Power Consumption" (IHEPC) (Hebrail & Berard, 2012), "Electricity Load Diagrams" (ELD) (Trindade, 2015), and "ENergy LIteracy Through an intelligent home ENergy advisor" (ENLITEN) (Lovett et al., 2016), which represent different electricity scenarios in order of high-dimensional, multivariate, and data-scarce, demonstrate the superior performance of our method and the promising application prospects of large models in ETSD anomaly detection problems.

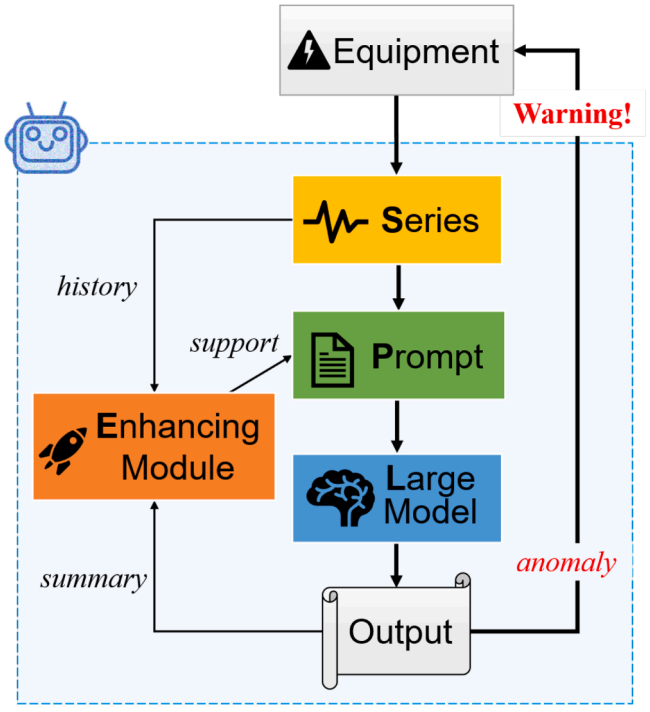

Fig. 1. Illustration of the SLEP smart agent framework.

We summarize the contributions of this paper as follows:

1. To the best of our knowledge, this paper is the first relevant work that leverages large models for time series anomaly detection in power systems. The SLEP agent framework is proposed as a novel common method to handle the series detection problem.
2. In order to allow the user to interact with the large model in a better way, we propose the new BRIDOR design template, which will considerably improve the quality of the prompt.
3. An implement of a smart agent based on the SLEP development framework is given and extensive experiments and cases have demonstrated the superior performance of it. Furthermore, we provide best practices for related problems, advancing the utilization of large models in the field of ETSD anomaly detection.

The remainder of the paper is organized as follows. Section 2 reviews research status of ETSD anomaly detection, large models and smart agents. Section 3 describes the SLEP framework and the BRIDOR template in detail. Section 4 presents an experimental implementation of the proposed method. Section 5 concludes this paper.

2. Related work

2.1. Anomaly detection in electric time series data

Currently, the methods and techniques (Cooper et al., 2023; Musleh et al., 2020; Susto et al., 2018; Zidi et al., 2023), for detecting anomalies in ETSD can be broadly categorized into two main groups: traditional hand-crafted methods and learning-based methods.

The traditional hand-crafted methods, encompassing statistical (Ho et al., 2020), model-based (Ashok et al., 2020), decomposition (Liu et al., 2021), and distance-based (Alencar et al., 2022) approaches, rely heavily on manual rule design, specific assumptions about data distribution, and often require clean, representative normal datasets. They are generally inflexible, struggle to adapt to the complexity and dynamics of real-world ETSD scenarios, and exhibit poor scalability. Due to these significant limitations, their popularity and applicability have waned considerably.

Accelerated by machine learning advancements, these data-driven approaches aim to automatically learn patterns and generalize to new data, offering potential improvements in accuracy and efficiency over traditional methods. Specifically, supervised learning methods (Habbak et al., 2023; Jindal et al., 2016; Shabad et al., 2021), such as SVM, Random Forest, and Decision Trees, frame anomaly detection as a binary classification task. While effective, they are fundamentally constrained by their dependence on large volumes of accurately labeled training data, which is notoriously scarce and costly to obtain in the electricity domain. This data scarcity severely limits their practical applicability. Unsupervised Learning Methods (Ahmed et al., 2019; Ceci et al., 2020; Gholami et al., 2024; Zanetti et al., 2019), including clustering, local outlier factor, and isolation forest, identify anomalies based on data density, distance, or isolation without needing labels. Although bypassing the labeling burden, their detection accuracy often falls short of

the high-reliability standards required in power systems. Furthermore, understanding why a specific point is flagged as an anomaly remains a significant challenge with these methods. Deep Learning Methods (Alshehri et al., 2024; Takiddin et al., 2023, 2022; Zhang et al., 2022), utilizing architectures like RNNs, LSTMs, and Autoencoders, excel at learning complex temporal dependencies and automatically extracting features. Anomalies are detected based on prediction errors or reconstruction losses. However, their powerful learning capability comes at the cost of requiring substantial amounts of data for training. Crucially, these complex models often function as "black boxes", offering limited insights into the underlying reasons for their anomaly predictions, which is critical for actionable diagnostics and trust in power system operations. In summary, while learning-based methods reduce the subjectivity of hand-crafted approaches, their advancement in ETSD anomaly detection is critically hampered by two intertwined challenges: (1) pervasive data scarcity impeding supervised and deep learning models, and (2) the lack of clear, human-understandable explanations for detected anomalies, particularly prominent in unsupervised and deep learning approaches. This highlights the urgent need for solutions that are both data-efficient and inherently interpretable.

To address these dual challenges, leveraging pre-trained large models emerges as a promising direction, which trained on massive and diverse datasets, offer the potential for effective knowledge transfer to downstream ETSD tasks, mitigating the impact of domain-specific data scarcity. Moreover, their advanced natural language understanding and generation capabilities provide a unique pathway to deliver intuitive, human-readable explanations for anomaly detections, significantly enhancing interpretability and trustworthiness in critical power system applications.

2.2. Large model and smart agent

In recent years, large models have emerged as a significant technological breakthrough in the field of artificial intelligence. These models are renowned for their complex network structures and immense parameter scales (Hadi et al., 2023; Min et al., 2023; Zhao et al., 2023). Examples include GPT, Bert, T5, LLaMA, and GLM, all of which possess billions, tens of billions, or even more parameters. With remarkable expressiveness and adaptability, large models have the ability to effectively understand and process natural language data, generating high-quality textual responses based on the given contextual descriptions. Typical large models are built upon the Transformer architecture (Vaswani et al., 2017). The early GPT-1 (Radford et al., 2018) was developed based on a generative, decoder-only Transformer. Subsequent GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and GPT-3.5 (Ouyang et al., 2022), adopted similar architectures and gradually extended the model's capabilities by continuously extending its parameter scale. The latest version in GPT family, GPT-4 (Achiam et al., 2023), has demonstrated astonishing generative abilities and cross-task performance, matching or even exceeding human-level performance on various professional and academic benchmarks. BERT (Devlin et al., 2018), another iconic large model, utilizes a bi-directional encoder Transformer architecture, leading to significant performance improvements on multiple natural language processing benchmarks. BERT and its successors, RoBERTa (Liu et al., 2019) and ERNIE (Sun et al., 2020), are widely employed in tasks such as text classification, named entity recognition, and question-answering systems. The T5 model (Raffel et al., 2020) also employs Transformer and achieves impressive performance across different scales of tasks through the introduction of adaptive scaling mechanisms. Meanwhile, emerging open-source large models like LLaMA and GLM have demonstrated strong competitiveness. LLaMA (Touvron et al., 2023) optimizes the Transformer architecture using techniques such as pre-normalization, SwiGLU activation function, and rotary embedding, and achieves excellent performance on public benchmark tests at relatively low computational cost. GLM (Du et al., 2022) surpasses state-of-the-art models like BERT and T5 in text understanding tasks through a self-regressive blank-filling Transformer architecture.

Although the large model has remarkable expressiveness and broad adaptability, it cannot be directly applied to ETSD anomaly detection. It is necessary to construct a smart agent (Ruan et al., 2023; Wang et al., 2024; Xi et al., 2025) to introduce the large model into the power system. By possessing the ability to perceive their surroundings, make decisions, and execute actions, agents could assist humans in achieving designated tasks. They have long been a focal point of attention for scholars in the field of artificial intelligence. At present, agent-led large models, due to their robust comprehension and generative capabilities, have demonstrated considerable potential for applications in a range of fields (Xue et al., 2024, 2025), including social sciences, natural sciences, engineering and so on.

To the best of our knowledge, there is currently no relevant work on large models for ETSD detection. Our research represents the first introduction of large models into this field.

3. Methodology

3.1. Problem description

We first give the time series anomaly detection problem definition. Given a multivariate time series set M as follows:

$$M = \left(x_1^1, x_1^2, \ldots, x_1^d\right), \left(x_2^1, x_2^2, \ldots, x_2^d\right), \ldots, \left(x_n^1, x_n^2, \ldots, x_n^d\right) \tag{1}$$

where each data point $X_t = (x_t^1, x_t^2, \ldots, x_t^d)$ is a vector of d variables that represents the observations at a given time point t. The goal of the problem of time series anomaly detection is to identify data points or subsets of data that exhibit either a failure to conform to an expected pattern or a significant deviation from normal behavior (anomaly) on more than one variable, either simultaneously or separately. That is, to find a multivariate anomaly detection function G that accepts as input a multivariate time series set M and outputs a set of anomalies $A \subseteq M$, where A contains the data points that are considered anomalous.

Specific to time series in electric scenarios, false data injection (FDI) methods are often used to introduce artificial anomalies for evaluating the performance of anomaly detection systems. Below, we introduce several prevalent FDI types (Zanetti et al., 2019; Zidi et al., 2023) and their formulations.

Constant subtraction: A fixed value a is subtracted from the selected data point X_t. The modified $\tilde{X}_t$ is calculated as

$$\tilde{X}_t = \max(X_t - a, 0) \tag{2}$$

where the lower bound 0 is enforced to avoid non-meaningful negative values.

Threshold-based clipping: Data point exceeding a predefined threshold is clipped. This FDI method is useful for concealing high-value anomalies (eg. high electricity consumption). The transformation function is given by

$$\tilde{X}_t = \begin{cases} X_t, & X_t < c \\ \tilde{c}, & X_t \geq c \end{cases} \tag{3}$$

where c is a randomly selected cut-off point, and $\tilde{c} \leq c$.

Zero-replacement: Data point at a randomly determined time period is set to zero. This type of FDI disrupts the normal data pattern, giving the appearance of no data occurrence during specific time. Mathematically, it is represented as

$$\tilde{X}_t = 0 \tag{4}$$

Proportional reduction: Data point is uniformly decreased by a constant percentage α, which is designed to consistently report data values lower than the actual ones. Mathematically, this can be expressed as

$$\tilde{X}_t = \alpha \cdot X_t, \quad 0 < \alpha < 1 \tag{5}$$

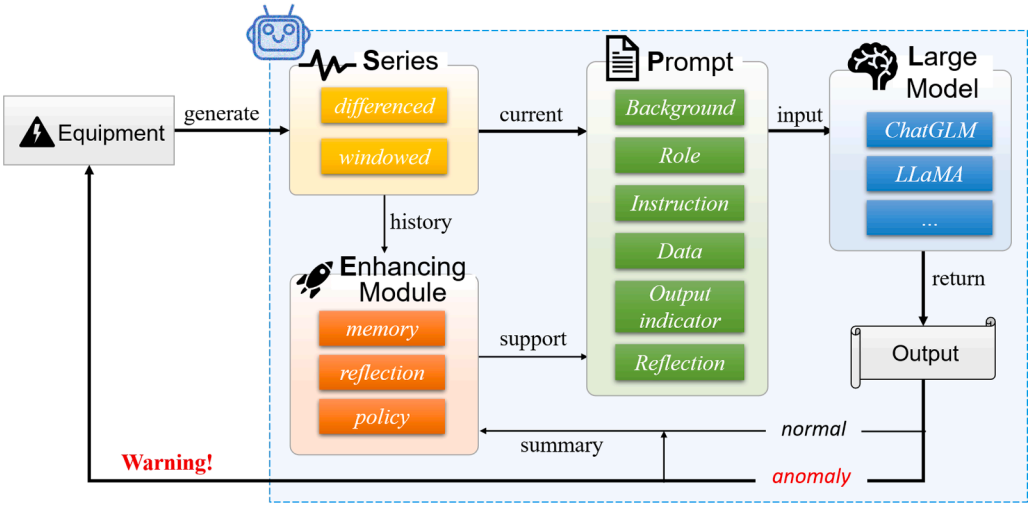

Fig. 2. Architecture of the large model-based smart agent development framework SLEP.

Specific percentage modification: Each data point is modified by a different randomly chosen percentage. The modified report is given by

$$\tilde{X}_t = \theta_t \cdot X_t, \quad 0 < \theta_t < 1 \tag{6}$$

where θ_t is randomly determined for each individual data point. This creates a more variable and less detectable form of FDI.

Historical average: Generates synthetic data points by multiplying the average value of the previous time period $\bar{X}_{\text{his}}$ by a randomly assigned percentage θ_t. This FDI method manipulates historical data to fabricate false data points and is represented as:

$$\tilde{X}_t = \theta_t \cdot \bar{X}_{\text{his}}, \quad 0 < \theta_t < 1 \tag{7}$$

3.2. The SLEP framework

Since the large model is merely an information processor and lacks the capacity for memory, perception, and action, it is necessary to construct a smart agent to fulfill the potential of the large model to handle the specified ETSD anomaly detection problem. However, none of the existing agents are oriented towards ETSD. Accordingly, we propose the large-model-based smart agent development framework SLEP (Fig. 2). It is tailored to the series detection problem and boasts quick deployment in any grid scenario that generates ETSD.

3.2.1. S for series

The "S" in SLEP refers to the time Series generated in real-time by the power equipment. In order to enhance the suitability of the model for larger datasets, preliminary pre-processing such as differencing and windowing are performed. The technique of differencing could reflect trends in series, reduce the effect of periodicity and the amount of data transmission, thus making it easier for the model to identify patterns or anomalies in the data. The differencing formula employed are recursively defined as follows:

$$\Delta^{(p)} x_i^j = \Delta^{(p-1)} x_{i+1}^j - \Delta^{(p-1)} x_i^j, \quad i = 1, 2, \ldots, n-p \tag{8}$$

where x is a data point in the series, totaling n; j denotes the dimension of x, being an integer taking values from 1 to d. Let p denote the difference dimension, being an integer taking values from 1 to $n-1$. When p is 1, $\Delta^{(p-1)} x = \Delta^{(0)} x = x$.

Another pre-processing approach in the S-module is windowing. The frequencies of detection and the quantities of ETSD generated by the involved power devices vary according to the specific power scenario. Direct process of these series may lead to insufficient computational resources or inefficient analysis. In this manner, the windowing technique is employed to split (combine) the ETSD into more manageable segments, thereby improving the detection efficiency. The form of a windowed series M_w is as follows:

$$M_w = \left(x_i^1, x_i^2, \ldots, x_i^d\right), \left(x_{i+1}^1, x_{i+1}^2, \ldots, x_{i+1}^d\right), \ldots, \left(x_{i+w-1}^1, x_{i+w-1}^2, \cdots, x_{i+w-1}^d\right) \tag{9}$$

where x is a data point in the series and w is the size of the window. Denote n as the total number of data points in the current series. When $w \geq n$, we have $w := n$; when $w < n$, set a step size s, let the starting position of the window $i = 1$, then cyclically slide s along the series (the new starting position $i' = i + s$), until the end of sliding.

3.2.2. L for large model

The "L" in SLEP represents the pre-trained Large models like ChatGLM and LLaMA. It serves as the "brain" of the smart agent, responsible for undertaking the key work of information judgement and analysis. In the series detection problem, the L-module is capable of automatically learning the patterns in the input ETSD data, determining whether there are any anomalous values and providing feedback.

3.2.3. E for enhancing

With the objective of enhancing the hand-crafted designed prompt through continuous evolution, we design the "E" Enhancing module within the framework. This module is structured as a self-adaptive system to facilitate better interaction with large models while continuously adapting to new challenges and data patterns. As shown in Fig. 3, the interior of Enhancing module can be divided into Series Memory, Reflection Library, and Enhancing Policy, details of which are described as follows:

Series Memory. Large models are stateless by default, meaning that each incoming query is processed independently, regardless of previous interactions. For these stateless large models, the only thing that exists is the current input and nothing else. However, in our series anomaly detection scenario, it is important to remember previous interactions, or the historical information of the power system. Therefore, we employed the series memory to store the historical series generated by the power equipment. Moreover, the memory will continuously update through adaptive forgetting mechanisms, based on the large model's suggestions stored in the reflection library. This ensures that recent and relevant data patterns are prioritized, while outdated information is systematically pruned. Subsequently, an appropriate enhancing policy will introduce these historical data into the prompt,

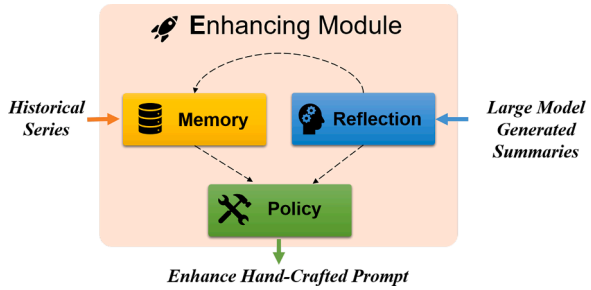

Fig. 3. Illustration of the enhancing module.

to enhance the relevance and accuracy of large models for problem processing.

Reflection Library. Reflection refers to the idea of letting large models check and optimize themselves. It is one of the most common patterns for designing intelligent agent workflows based on large models that can easily and quickly lead to performance improvements. By embedding reflection commands in prompts such as "please check the quality and accuracy of the answer you gave and its relevance to this prompt, and make constructive suggestions on how to improve the answer and the prompt", we can allow large models to evaluate their work and make suggestions for improvement. Constantly updated reflections will be recorded and stored in the reflection library, which will provide an important basis for iterative improvement of the prompts.

Enhancing Policy. Although general large models are powerful, they are prone to difficulties in areas where traditional computers excel, such as logic, computation, and search. Anomaly detection in ETSD is a time-honored problem that has been explored and researched extensively by many scholars. The ideas embedded in all these works can be promising as enhancing policies to provide additional knowledge and information to the prompt. In the SLEP framework, we have provided three typical enhancing policies as optional parameters for the developer, i.e., statistical information, features and a past period of historical series, to adapt different requirements in grid scenarios.

3.2.4. P for prompt

"P" stands for Prompt. It is the most critical interaction module in the SLEP framework, linking the S and L modules. Prompt is the guiding text provided by the user (outside system) to interact with the large model, which clarifies the logics and goals of the problem and steers the behavior of the smart agent. More precise prompt words can enable the large models to produce more demand-responsive results. To improve the quality of prompt words, we propose the BRIDOR design template. (Detailed in Section 3.3)

3.3. BRIDOR template for prompt

Prompts serve as the interface through which the external world engages with large models. They act as the bridge that translates user intentions and requirements into instructions that large models can understand and process. With a mere alteration of the instructions within the prompt, the same large model has the flexibility to address a diverse array of tasks, demonstrating the adaptability and versatility of the large model and the pivotal impact of prompts on performance. In fact, the quality of prompts is a decisive factor in the effectiveness of leveraging general large models to accomplish specific tasks. Well-crafted prompts could guide large models to generate accurate outputs, while poor ones lead to incorrect results and affect task performance. To address the challenge of designing effective prompts, we studied the key components that contribute to high-quality prompts and developed the BRIDOR design template. It offers a standardized, structured, and systematic pattern for the construction of prompts. Comprising the following six essential components (Fig. 4), the BRIDOR template serves as a comprehensive guide for creating prompts that can effectively communicate with large models and elicit optimal responses.

Background (B): This part provides large models with the necessary external knowledge and contextual information, which helps large models to understand the task requirements more accurately and generate the corresponding responses. It can be edited manually, retrieved from external databases, or introduced into the prompt through plug-ins and calculations. In the SLEP framework, this is populated by the Enhancing module.

Role (R): It clarifies the specific role that large models play in the interaction and can mobilize the relevant knowledge within the large models to generate a more targeted and professional output. For our problem, the role large models need to play is that of a power system engineer with seasoned experience.

Instruction (I): This part informs large models of the specific tasks they need to perform, including what to do, how to do it, how to use the background information provided (if any), how to handle the input data, etc. To do the detection, this part needs to explain to the large models the problem described in the Section 3.1.

Input data (D): It is the specific data that the user needs large models to process. In our application, it is the current time series M to be checked for anomalies.

Output indicator (O): This part is used to set the format of large model's outputs, so that it can be matched with other components for automated processing. For the anomaly detection problem, the output should be a comment starting with "Normal" or "Abnormal". (E.g., for normal time series, the output might be: "Normal, no abnormality was found in the given data.")

Reflection (R): It requires large models to self-assess after generating responses and optimize future interactions by reflecting on the quality, accuracy and usefulness of their responses. In our application

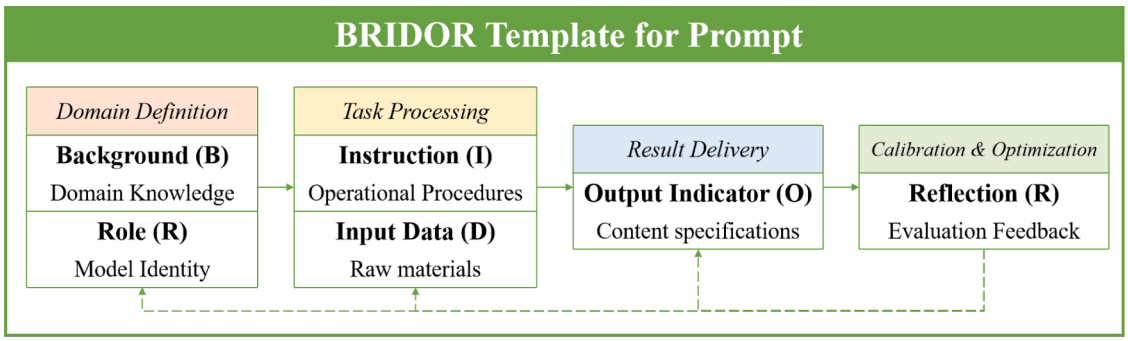

Fig. 4. Six components and their functions of the BRIDOR prompt template.

a) a simple spoken prompt

Here are some historical data for the series *[232.69, 230.98, 232.21, ...]*, and I'd like to ask you to see if there are any outliers in the subsequent *[236.66, 235.84, 235.6, ..., 233.62, 234.68]*? If so, can you point out where it is?

b) a relatively specific prompt

Here are some historical data for the series *[232.69, 230.98, 232.21, ...]*. Next, I will send you a series of data, 10 at a time, in chronological order of their generation. If you think these 10 data are normal, reply "normal" and predict the possible direction of the next, otherwise reply "abnormal", point out their locations, and explain the reason. *[236.66, 235.84, 235.6, ..., 233.62, 234.68]*.

c) a carefully constructed prompt based on BRIDOR

Background (B): *232.69, 230.98, 232.21, ...*
Role (R): You are a power system engineer who has been in the field for many years
Instruction (I): Now I need your help to detect the presence of outliers in a sequence of power time series. The data points in the sequence to be detected will be arranged in the chronological order of their generation, separated by ",", totaling 10. You can refer to the historical series provided in the background information to take advantage of its overall pattern, distribution, or other features that you think are important.
Input Data (D): In *[236.66, 235.84, 235.6, ..., 233.62, 234.68]* are the power time series that I need you to carry out to detect if there are outliers.
Output Indicator (O): For the specific judgement part of this series, your reply should be separated from the other parts, which include: if you think that the data points in this series are normal, in line with the direction of the series in your opinion, reply "[Judgement: Normal]", and try to predict the trend of the subsequent data. If you think that the sequence is abnormal, and some data points do not match the direction of the sequence in your opinion, reply "[Judgement Result: Abnormal]" and the location or data points that you think are abnormal. You will then need to specify the basis for your "normal" or "abnormal" inference, such as the overall pattern, distribution, or other characteristics of the sequence data that you think are important.
Reflection (R): Please scrutinize the quality and accuracy of the response you have given and its relevance to the prompt, and make constructive suggestions on how to improve the response and the prompt.

Fig. 5. Three examples of manually designed prompts. Red, green, and blue highlight the direct task instructions (including the expected replies), the historical series for reference and the current series to be detected.

scenario, we will use the reflection component at an early stage so that large models can adjust and refine the prompts, based on their previous responses and the feedback from the users (the power system), avoiding subjective interferences implied in the manually designed initial prompts.

These six interconnected components synergistically collaborate to construct a comprehensive and highly efficient prompt architecture, which could enhance the large model's comprehension of tasks, optimize its response generation process, and elevate the overall performance. Depending on the specific task requirements and application scenarios, a prompt may incorporate fewer than all the six components of BRIDOR. However, to ensure the effectiveness and reliability, a well-designed prompt typically encompasses at least two or more components. This flexibility allows for tailored prompt construction while maintaining the core principles of structured design.

To provide a more intuitive and tangible understanding of the BRIDOR template, Fig. 5 presents three distinct examples of manually designed prompts. The first example (a) is a simple, conversational prompt formulated solely based on the problem description, lacking a structured approach. The second example (b) demonstrates a more specialized prompt, designed with basic knowledge of prompt engineering, but still limited in its comprehensiveness. In contrast, the third example (c) showcases a meticulously crafted prompt constructed using the BRIDOR template. It can be seen that the BRIDOR-based prompt exhibits a higher degree of structure and systematic organization. This enhanced design not only facilitates a more profound understanding of tasks by the large model but also significantly improves its task execution capabilities, resulting in more accurate, relevant, and valuable responses. (An experimental validation was carried out in Section 4.3.2)

4. Experiments

4.1. Setup

4.1.1. Evaluation metrics

The detection result can be represented by a confusion matrix, as shown in Table 1. In this matrix, "TP" (True Positive) denotes the number of instances where actual anomalies are correctly identified by the detection model. "FN" (False Negative) represents the number of actual anomalies that the model fails to detect. "FP" (False Positive) refers to the cases where normal data is misclassified as anomalous. "TN" (True Negative) indicates the number of times normal data is correctly classified as normal. Each element plays a crucial role in evaluating the performance of the detection model

We chose the following commonly used metrics (Alshehri et al., 2024; Habbak et al., 2023; Zidi et al., 2023) to evaluate the performance of series anomaly detection.

Accuracy is defined as the overall accuracy of the detection, indicating the frequency with which it correctly identifies both normal and

Table 1
Detection confusion matrix. Anomaly treated as positive.

Detected \ Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 2
The detection performance (%) on IHEPC, with 4 valid digits retained and the optimal metrics highlighted in bold.

Large Model	Policy	Accuracy	Precision	Recall	F1
ChatGLM3-6B	B	99.07	81.13	*95.56*	87.76
	S	98.58	70.56	100.00	82.74
	F	98.04	83.33	*50.30*	62.73
	H	**99.86**	**96.46**	100.00	**98.20**
Llama2-7B	B	98.02	40.48	100.00	57.63
	S	98.04	36.94	100.00	53.95
	F	98.53	39.80	*72.22*	51.32
	H	**99.90**	**92.19**	100.00	**95.93**
Comparison Methods					
DSVDD (Habbak et al., 2023)		97.42	83.33	*42.08*	55.92
RETAD (Takiddin et al., 2023)		99.05	82.52	*92.89*	87.40
FC-AE (Alshehri et al., 2024)		99.38	88.78	*93.14*	90.90

Table 3
The detection performance (%) on ELD, with 4 valid digits retained and the optimal metrics highlighted in bold.

Large Model	Policy	Accuracy	Precision	Recall	F1
ChatGLM3-6B	B	98.28	46.95	100.00	63.90
	S	98.24	46.39	100.00	63.37
	F	98.61	54.00	*69.23*	60.67
	H	**99.88**	**92.50**	100.00	**96.10**
Llama2-7B	B	98.00	36.88	100.00	53.88
	S	98.71	48.80	*98.39*	65.24
	F	98.67	44.00	*80.00*	56.77
	H	**99.90**	**90.00**	100.00	**94.74**
Comparison Methods					
DSVDD (Habbak et al., 2023)		97.51	80.00	*42.66*	55.65
RETAD (Takiddin et al., 2023)		99.32	81.16	*98.25*	88.89
FC-AE (Alshehri et al., 2024)		99.58	92.65	*95.35*	93.98

anomalous values.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{10}$$

Precision evaluates the effectiveness in distinguishing true anomalies from the values it flags as potentially anomalous. Higher levels of precision indicate a lower false alarm rate.

$$precision = \frac{TP}{TP + FP}. \tag{11}$$

Recall, on the other hand, assesses the method's ability to detect all actual anomalies within the series. A higher level of recall indicates a lower percentage of underreporting. In order to comply with the high security requirements of the power system, it is vital that recall is at a high level, ensuring that no anomalies are missed.

$$recall = \frac{TP}{TP + FN}. \tag{12}$$

The F1 score represents a balanced indicator between precision and recall, offering a single measure that encapsulates both aspects.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{13}$$

All these metrics range from 0 to 1, the higher the better. And together, they provide a comprehensive evaluation of the detection's performance.

Table 4
The detection performance (%) on ENLITEN, with 4 valid digits retained and the optimal metrics highlighted in bold.

Large Model	Policy	Accuracy	Precision	Recall	F1
ChatGLM3-6B	B	98.02	38.89	*98.44*	55.75
	S	98.57	51.02	100.00	67.57
	F	98.69	52.00	*74.29*	61.18
	H	**99.80**	**86.84**	100.00	**92.96**
Llama2-7B	B	99.00	70.58	*99.17*	82.47
	S	99.01	72.67	100.00	84.17
	F	98.45	74.25	*58.59*	65.50
	H	**99.86**	**95.62**	100.00	**97.76**
Comparison Methods					
DSVDD (Habbak et al., 2023)		98.11	82.35	*51.53*	63.40
RETAD (Takiddin et al., 2023)		98.59	64.67	*99.23*	78.31
FC-AE (Alshehri et al., 2024)		99.02	78.16	*98.55*	87.18

4.1.2. Implementation details

The proposed method is implemented based on the LangChain (Topsakal & Akinci, 2023) software library, with ChatGLM3-6B (Du et al., 2022) and LLaMA2-7B (Touvron et al., 2023) chosen as the base large models. ChatGLM3-6B is an open-source model in the ChatGLM3 series of dialogue pre-training models. It has 6 billion parameters and can deal with a context length of up to 8K. LLaMA2-7B is the 7-billion-parameter version of the open-source Llama2 large model series, which can handle a context length of up to 4K. Both of them can reflect the strongest performance of the current open-source general large models.

For the ETSD employed, the first 30 % was used as priori historical data for statistics, training or recording, and the last 70 % was used for the detection testing experiments. Anomalies were randomly injected using the methods described in Section 3.1, with the injection rate set to $1e-3$,

The experimental hardware platform consists of an 8-core Intel(R) Xeon(R) Platinum 8369 CPU, a NVIDIA RTX 4090 GPU featuring 24 GB of graphics memory, 32 GB of DDR4 RAM, and operates on Ubuntu 22.04.

4.2. Case studies

To verify the effectiveness of the proposed method, we use three real electricity datasets, IHEPC (Hebrail & Berard, 2012), ELD (Trindade, 2015) and ENLITEN (Lovett et al., 2016) to conduct the experiments. The three real-world cases are from France, Portugal, and the United Kingdom respectively, each exemplifying distinct electricity scenarios, in the order of high-dimensional, multivariate, and data-scarce. More details and evaluation results are provided in the following sections.

Our method was carried out using the proposed SLEP development framework, with prompts designed according to our BRIDOR template, and the window size set to 50. We compared the performance of four different policies: basic prompt (B), prompt enhanced with statistical information (S), prompt enhanced with learning features (F), and prompt enhanced with historical series (H).

Three methods DSVDD (Habbak et al., 2023), RETAD (Takiddin et al., 2023) and FC-AE (Alshehri et al., 2024) were chosen for the performance comparison, all of which are well-performing learning-based ETSD detection methods in recent years. According to the original paper, DSVDD needs to convert the time series into an image first, and we used a 5×10 grayscale image to meet the window size 50. The maximum training epoch for both methods was set to 100 and all codes used were our own production.

4.2.1. Real-world case 1

The IHEPC (Hebrail & Berard, 2012) dataset contains $2,075,259$ records of electricity consumption data gathered in a house located in Sceaux (a small town near Paris, France) over a 47-month period

Table 5
The detection performance (%) of different window sizes, with 4 valid digits retained and the optimal metrics highlighted in bold.

Dataset	Window Size	ChatGLM3-6B				Llama2-7B			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
IHEPC	10	98.96	77.02	100.00	87.02	99.15	52.22	100.00	68.61
	30	99.75	93.12	100.00	96.44	99.86	87.93	100.00	93.58
	50	99.86	96.46	100.00	98.20	99.90	92.19	100.00	95.93
	80	**99.88**	**96.67**	100.00	**98.31**	99.92	93.85	100.00	96.83
	100	99.84	95.72	100.00	97.81	**100.00**	**100.00**	**100.00**	**100.00**
ELD	10	99.08	66.19	100.00	79.65	99.15	51.69	100.00	68.15
	30	99.82	90.32	100.00	94.92	99.86	89.06	100.00	94.21
	50	99.88	92.50	100.00	96.10	99.90	90.00	100.00	94.74
	80	**99.96**	**97.06**	100.00	**98.51**	**99.98**	**98.11**	100.00	**99.05**
	100	99.94	96.30	100.00	98.11	**99.98**	97.67	100.00	98.82
ENLITEN	10	99.13	58.49	100.00	73.81	99.06	73.03	100.00	84.42
	30	99.82	90.00	100.00	94.74	99.75	90.78	100.00	95.17
	50	99.80	86.84	100.00	92.96	99.86	95.62	100.00	97.76
	80	99.92	95.00	100.00	97.44	99.94	97.52	100.00	98.74
	100	**99.94**	**96.59**	100.00	**98.27**	**99.96**	**99.32**	*99.32*	**99.32**

Table 6
The average time consumption (s) for one detection of ChatGLM3-6B under different window sizes, on ENLITEN. Note: to avoid additional time consumption, large model reflection is disabled here.

Window Size	10	30	50	80	100
Time Consumption	0.63	0.65	0.78	0.92	1.14

between December 2006 and November 2010. The sampling rate is one-minute. Attributes such as time, active power, reactive power, voltage and sub-meter readings are available, which effectively characterize the high-dimensional scenario.

Results conducted on IHEPC are presented in Table 2, with 4 valid digits retained and the optimal metrics highlighted in bold. As can be seen from the table, the *accuracy* of all four policies of our method reach over 98.02 %, with the optimal case even reaching 99.90 %. This is on the level of the comparison methods. In terms of *recall*, which is crucial for anomaly detection, more than half of the policies achieve 100 %. This exceeds the comparison methods. Fig. 6 gives the visualization comparison results conducted on the IHEPC dataset.

4.2.2. Real-world case 2

The ELD (Trindade, 2015) dataset consists 140,256 consumption records for 370 electrical loads in Portugal from 2011 to 2014. (Some loads were gathered after 2011, whose previous measurements were set to 0.) The sampling rate is 15-min, yielding 96 (24×4) measurement values a day per load. Each load measurement is a time-series variable, collectively representing a good multivariate scenario.

Results conducted on ELD are presented in Table 3, with 4 valid digits retained and the optimal metrics highlighted in bold. It can be seen that the metric *accuracy* of all four policies of our method are equal or over 98.00 %, while the best case is reaching 99.90 %. This is comparable to the two comparison methods. And regarding the crucial metric *recall*, more than half of our policies achieve 100 %. This means that our method is able to find all anomalies, which is superior to the comparison methods. Fig. 7 gives the visualization comparison results conducted on the ELD dataset.

4.2.3. Real-world case 3

The ENLITEN (Lovett et al., 2016) project was deployed in 200 homes for a period of 2 years in Exeter (the county seat of Devon, UK). The publicly available ENLITEN dataset consists sensor data from August-October 2013 for a smart houses, with attributes including electricity consumption of various appliances, light levels, carbon dioxide

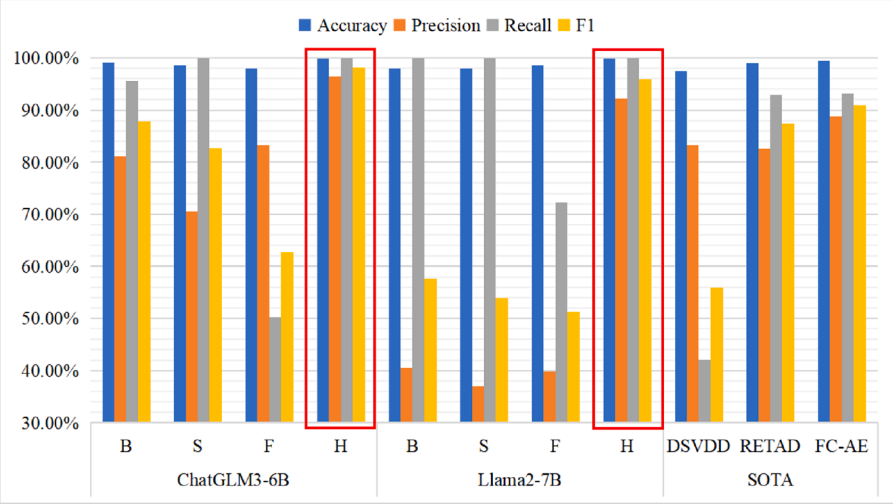

Fig. 6. Visualization results on IHEPC. The best results are boxed up.

Table 7
The detection performance (%) of different prompts, with 4 valid digits retained and the optimal metrics highlighted in bold.

Dataset	Prompt	ChatGLM3-6B				Llama2-7B			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
IHEPC	a	99.44	85.99	100.00	92.47	99.80	83.05	100.00	90.74
	b	**99.92**	95.56	100.00	97.73	**99.90**	90.00	100.00	94.74
	c	99.86	**96.46**	100.00	**98.20**	**99.90**	**92.19**	100.00	**95.93**
ELD	a	99.59	81.74	100.00	89.95	99.66	77.63	100.00	87.41
	b	**99.90**	90.91	100.00	95.24	99.86	87.50	100.00	93.33
	c	99.88	**92.50**	100.00	**96.10**	**99.90**	**90.00**	100.00	**94.74**
ENLITEN	a	99.59	79.81	100.00	88.77	99.51	85.29	100.00	92.06
	b	**99.82**	85.25	100.00	92.04	99.80	92.31	100.00	96.00
	c	99.80	**86.84**	100.00	**92.96**	**99.86**	**95.62**	100.00	**97.76**

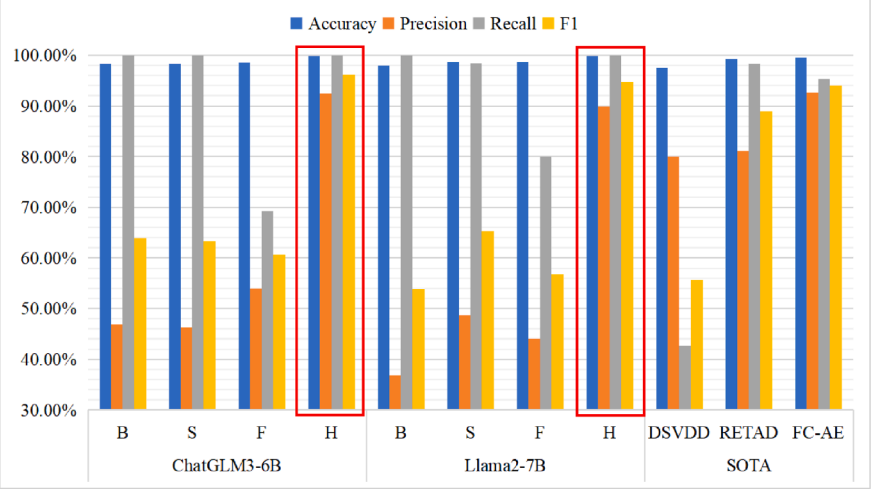

Fig. 7. Visualization results on ELD. The best results are boxed up.

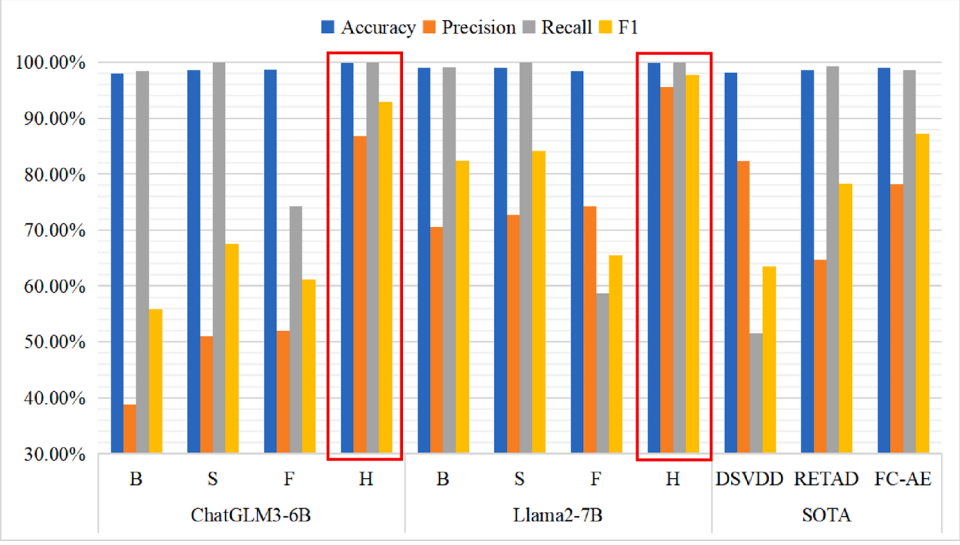

Fig. 8. Visualization results on ENLITEN. The best results are boxed up.

concentration, etc. The experiments in this paper employed only the electricity-related data contained therein. Due to the relatively small scope and short time for its data collection, ENLITEN serves as a case for the data-scarce scenario.

Results conducted on ENLITEN are presented in Table 4, with 4 valid digits retained and the optimal metrics highlighted in bold. In the table, it appears that the *accuracy* of all four policies of our method are equal or over 98.02 %, while the best case is reaching 99.86 %. This is paralleled with the comparison methods. And for the *recall* rate, a crucial anomaly detection metric, half of our policies achieve 100 %. This outperforms the comparison methods. Fig. 8 gives the visualization comparison results conducted on the ENLITEN dataset.

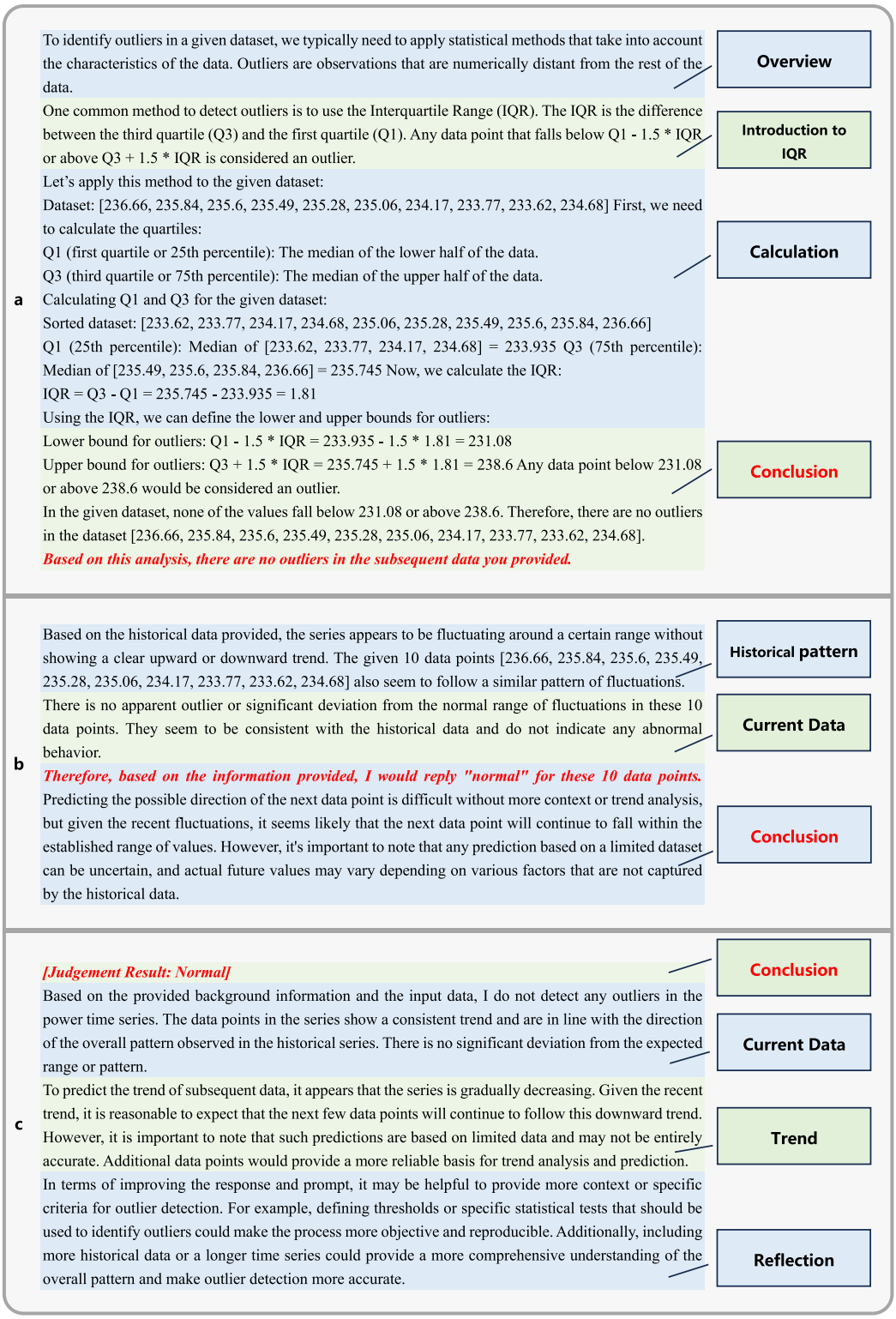

Fig. 9. Replies given by the large model using the three prompts shown in Fig. 5. Mark up according to the general idea and highlight the key conclusions.

4.2.4. Discussion

As can be seen from the cases, all policies of our method can reach over an *accuracy* of 98 %. And more than half of them could reach a 100 % *recall* rate, indicating that these approaches can identify all anomalies in the power system and issue maintenance alerts promptly. This demonstrates that our large model-based methods possess the capability to detect anomalies in different power scenarios.

Among the four policies we chose, policy H (enhance prompt by historical series) achieved the best performance in all metrics (it beats the two comparison methods as well). This is perhaps because it provides only objective historical information, avoiding potential biases that may be introduced by manually selected statistical information (such as maximum, minimum, quartiles, average, variance, etc.) and pre-trained learning models. Thus, it can better mobilize the potential

Table 8
The average time consumption (s) for one detection of ChatGLM3-6B under different prompts, on ENLITEN. Note: to avoid additional time consumption, large model reflection is disabled here.

Prompt	a	b	c
Time Consumption	1.75	0.91	0.78

knowledge contained in large models. These practices highly inspire our future researches.

4.3. Further analysis

4.3.1. Analysis of different window sizes

We further analyze the impact of different window sizes based on the well-behaved policy H in Section 4.2. The experimental results are shown in Tables 5 and 6.

It can be seen that the metrics increase as the window size increases. A full 100 % detection performance was even obtained on the IHEPC dataset using a window size of 100. This suggests that providing relatively more information benefits the large model in its judgement (which is consistent with the reflection by large models). However, series are time-sensitive, and as shown in Table 6, with ChatGLM3-6B on ENLITEN as an instance, the inference of large models takes time, larger window will lag the detection. It is therefore necessary to choose the window size appropriately according to the actual scenario.

4.3.2. Analysis of different prompts

When introducing the BRIDOR template in Section 3.3, three examples of different prompts were given. Here we will analyze more. Policy H and a window size of 50 are chosen for the experiments. Results are shown in Table 7. It is easy to see that the prompt c designed based on our BRIDOR template achieves better performance.

Furthermore, we present the average time consumption under the three different styles prompt in Table 8, with ChatGLM3-6B on ENLITEN as the instance. A casual prompt (a) causes the large model to engage in more autonomous comprehension, increasing its inference time; an organized prompt (c) clarifies task requirements with precision, leading to a reduction in processing duration; and the in-between prompt (b) exhibits intermediate elapsed time. This empirical evidence underscores that interacting with large models through structured prompts significantly enhances inference efficiency. Notably, the proposed BRIDOR template provides a systematic way to construct such structured prompts, optimizing human-model interaction and reducing time costs.

Fig. 9 shows the cases of replies given by the large model using the three prompts. All 3 cases correctly determine whether the input series is anomalous or not. Among them, reply a includes a large number of intermediate processes, which not only consumes time but also hinders the rapid parsing by automated functions. Reply b is relatively brief, yet its output format is not fixed and lacks a controllable reflection part. In contrast, reply c (derived from the BRIDOR-based prompt) is concise and easy to automate. This qualitatively suggests that the structured prompt construction template BRIDOR can effectively optimize the large model's understanding and reasoning processes, while reducing the interaction difficulty with other automated modules within our SLEP smart agent framework.

5. Conclusion

In this paper, we design a large-model-based automated agent method for ETSD anomaly detection, and show the excellent performance of the proposed method through extensive experiments conducted on the high-dimensional, multivariate, and data-sparse real-world cases, i.e., IHEPC, ELD, and ENLITEN datasets. These practices amply demonstrate that knowledge-enriched large models can be leveraged to quickly solve many power system problems in completely different scenarios, thus greatly improving efficiency and reducing labor costs. However, the huge scale of large models leads to higher storage costs and longer reasoning time, which will somewhat limit the application of the proposed method in some practical scenarios. Therefore, our future research direction will be the lightweight deployment of large models for specific tasks, as well as more types of anomalies and richer industrial application scenarios, to promote the continuous development of related technologies.

CRediT authorship contribution statement

Bingrui Wang: Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Yuan Zhou:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing; **Leijiao Ge:** Writing – review & editing; **Sun-Yuan Kung:** Supervision, Writing – review & editing.

Data availability

Data will be made available on request.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62171320).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ahmed, S., Lee, Y., Hyun, S.-H., & Koo, I. (2019). Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest. *IEEE Transactions on Information Forensics and Security*, *14*(10), 2765–2777.

Alencar, G. T., Santos, R. C., & Neves, A. (2022). Euclidean distance-based method for fault detection and classification in transmission lines. *Journal of Control, Automation and Electrical Systems*, *33*(5), 1466–1476.

Alshehri, A., Badr, M. M., Baza, M., & Alshahrani, H. (2024). Deep anomaly detection framework utilizing federated learning for electricity theft zero-day cyberattacks. *Sensors*, *24*(10), 3236.

Ashok, A., Govindarasu, M., & Ajjarapu, V. (2020). Model-based anomaly detection for power system state estimation. In *Advances in Electric Power and Energy: Static State Estimation* (pp. 99–121).

Bhattarai, B. P., Paudyal, S., Luo, Y., Mohanpurkar, M., Cheung, K., Tonkoski, R., Hovsapian, R., Myers, K. S., Zhang, R., Zhao, P. et al. (2019). Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid*, *2*(2), 141–154.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Ceci, M., Corizzo, R., Japkowicz, N., Mignone, P., & Pio, G. (2020). ECHAD: Embedding-based change detection from multivariate time series in smart grids. *IEEE Access*, *8*, 156053–156066.

Cooper, A., Bretas, A., & Meyn, S. (2023). Anomaly detection in power system state estimation: Review and new directions. *Energies*, *16*(18), 6678.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, F., Chen, S., Demachi, K., Yoshikawa, M., Seki, A., & Takaya, S. (2023). Attention-based time series analysis for data-driven anomaly detection in nuclear power plants. *Nuclear Engineering and Design*, *404*, 112161.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022). GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 320–335).

Gholami, A., Tiwari, A., Qin, C., Pannala, S., Srivastava, A. K., Sharma, R., Pandey, S., & Rahmatian, F. (2024). Detection and classification of anomalies in power distribution system using outlier filtered weighted least square. *IEEE Transactions on Industrial Informatics*, *20*(5), 7513–7523.

Habbak, H., Mahmoud, M., Fouda, M. M., Alsabaan, M., Mattar, A., Salama, G. I., & Metwally, K. (2023). Efficient one-class false data detector based on deep SVDD for smart grids. *Energies*, *16*(20), 7069.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S. et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints.

Hebrail, G., & Berard, A. (2012). Individual household electric power consumption. UCI Machine Learning Repository. https://doi.org/10.24432/C58K54

Ho, C. H., Wu, H. C., Chan, S. C., & Hou, Y. (2020). A robust statistical approach to distributed power system state estimation with bad data. *IEEE Transactions on Smart Grid*, *11*(1), 517–527.

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, *12*(3), 1005–1016.

Kataray, T., Nitesh, B., Yarram, B., Sinha, S., Cuce, E., Shaik, S., Vigneshwaran, P., & Roy, A. (2023). Integration of smart grid with renewable energy sources: Opportunities and challenges–a comprehensive review. *Sustainable Energy Technologies and Assessments*, *58*, 103363.

Li, H., Liu, Z., Chen, X., Yuan, W., Kaleem, M. B., & Liu, W. (2023). Early anomaly detection of power battery based on time-series features. In *2023 3rd New energy and energy storage system control summit forum (NEESSC)* (pp. 383–388). IEEE.

Liu, J., Wu, S., Cao, W., Guo, Y., & Gong, S. (2021). Smart grid data anomaly detection method based on cloud computing platform. In *Artificial intelligence and security: 7th International conference, ICAIS 2021, Dublin, Ireland, July 19–23, 2021, proceedings, Part I 7* (pp. 338–345). Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Lovett, T., Lee, J., Gabe-Thomas, E., Natarajan, S., Brown, M., Padget, J., & Coley, D. (2016). Designing sensor sets for capturing energy events in buildings. *Building and Environment*, *110*, 11–22. https://doi.org/10.1016/j.buildenv.2016.09.004

Luan, J. (2023). ChatGPT+ vector database + prompt-as-code - the CVP stack. https://zilliz.com/blog/ChatGPT-VectorDB-Prompt-as-code.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

Musleh, A. S., Chen, G., & Dong, Z. Y. (2020). A survey on the detection algorithms for false data injection attacks in smart grids. *IEEE Transactions on Smart Grid*, *11*(3), 2218–2234. https://doi.org/10.1109/TSG.2019.2949998

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018). Improving language understanding by generative pre-training. OpenAI blog.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., qing, d., shiwei, s., Mao, H., Zeng, X., & Zhao, R. (2023). TPTU: Task planning and tool usage of large language model-based AI agents. In *NeurIPS 2023 foundation models for decision making workshop*. https://openreview.net/forum?id=GrkgKtOjaH.

Shabad, P. K. R., Alrashide, A., & Mohammed, O. (2021). Anomaly detection in smart grids using machine learning. In *IECON 2021–47th Annual conference of the IEEE industrial electronics society* (pp. 1–8). IEEE.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8968–8975). (*vol. 34*).

Susto, G. A., Cenedese, A., & Terzi, M. (2018). Time-series classification methods: Review and applications to power systems data. In *Big Data Application in Power Systems* (pp. 179–220).

Takiddin, A., Ismail, M., & Serpedin, E. (2023). Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids. *IEEE Transactions on Smart Grid*, *14*(1), 663–676. https://doi.org/10.1109/TSG.2022.3193989

Takiddin, A., Ismail, M., Zafar, U., & Serpedin, E. (2022). Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids. *IEEE Systems Journal*, *16*(3), 4106–4117.

Topsakal, O., & Akinci, T. C. (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences* (pp. 1050–1056). (*vol. 1*).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023). LLAMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Trindade, A. (2015). ElectricityLoadDiagrams20112014. UCI Machine Learning Repository. https://doi.org/10.24432/C58C86

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 6000–6010.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y. et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 1–26.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E. et al. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), 121101.

Xue, Y., Jin, G., Shen, T., Tan, L., Wang, N., Gao, J., & Wang, L. (2024). Consistent representation mining for multi-drone single object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(11), 10845–10859. https://doi.org/10.1109/TCSVT.2024.3411301

Xue, Y., Zhong, B., Jin, G., Shen, T., Tan, L., Li, N., & Zheng, Y. (2025). AVLTrack: Dynamic sparse learning for aerial vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2025.3549953

Zanetti, M., Jamhour, E., Pellenz, M., Penna, M., Zambenedetti, V., & Chueiri, I. (2019). A tunable fraud detection system for advanced metering infrastructure using short-lived patterns. *IEEE Transactions on Smart Grid*, *10*(1), 830–840.

Zhang, G., Li, J., Bamisile, O., Cai, D., Hu, W., & Huang, Q. (2022). Spatio-temporal correlation-based false data injection attack detection using deep convolutional neural network. *IEEE Transactions on Smart Grid*, *13*(1), 750–761. https://doi.org/10.1109/TSG.2021.3109628

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Zidi, S., Mihoub, A., Qaisar, S. M., Krichen, M., & Al-Haija, Q. A. (2023). Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment. *Journal of King Saud University-Computer and Information Sciences*, *35*(1), 13–25.