Zero-shot GOOSE Anomaly Detection via Multi-gate Mixture-of-Experts with Pre-Trained Large Language Model

Yi Li*, Mingfeng Fan*, Guo Chen†, Chaojie Li†, and Biplab Sikdar*
*Department of Electrical and Computer Engineering, National University of Singapore
†School of Electrical Engineering and Telecommunications, University of New South Wales
Email: liyi_42@u.nus.edu, {mingfeng.fan, bsikdar}@nus.edu.sg, guo.chen@unsw.edu.au, chaojie.li2019@gmail.com

***Abstract*—Detecting anomalies in smart grids is vital to safeguarding systems from attacks and failures. As critical components in IEC 61850-based substation communication, Generic Object-Oriented Substation Event (GOOSE) messages are particularly vulnerable to replay, insertion, and flooding attacks, which can compromise availability. However, existing anomaly detection methods mainly focus on traditional network flows like TCP/IP, neglecting the semantic information and structured characteristics of GOOSE messages. This limits the ability to exploit rich information and detect potential attack indicators. Moreover, imbalanced datasets and unseen anomaly types pose additional challenges, highlighting the need for robust few-shot and zero-shot learning approaches. To address these challenges, we propose GAMMPT framework for GOOSE anomaly detection. GAMMPT first leverages pre-trained large language models to extract semantic features and then tackles data imbalance by decomposing multi-class detection into binary classification tasks to improve precise anomaly type recognition. Subsequently, it employs an attention-based Multi-gate Mixture-of-Experts (attMMoE) model to enhance few-shot learning through shared experts and improve anomaly detection accuracy. To enhance zero-shot learning, GAMMPT clusters GOOSE messages and incorporates contrastive learning to enhance embedding robustness. Experiment shows that GAMMPT achieves state-of-the-art performance on real-world datasets.**

***Index Terms*—GOOSE Anomaly Detection, Zero-Shot, Cyber-Security, Multi-Gate Mixture-of-Experts, Large Language Model**

I. Introduction

Ensuring the optimal operation of power grids requires robust communication security for critical infrastructures [1]. In substation communication networks, data availability often takes precedence over confidentiality and integrity [2]. IEC 61850 is widely adopted as a communication standard for smart grids, which can facilitate seamless communication among intelligent electronic devices (IEDs) within substations. GOOSE (Generic Object Oriented Substation Event) messages serve as an important part of IEC 61850-based substation communication networks, enabling the efficient and real-time monitoring of substation operations. This functionality is essential for maintaining the safety, stability, and efficiency of the substation. However, since the IEC 61850 standard does not include built-in security features, GOOSE messages are vulnerable to replay, insertion, and flooding attacks, which can significantly compromise data availability and disrupt the substations. Hence, the cyber-security of GOOSE has been identified as a vital issue for the substations.

Recently, there has been growing interest within the research community in enhancing GOOSE cyber-security for modern smart grids. Several studies have explored encryption-based security measures for GOOSE messages [3], but failed to meet the time-criticality requirements. Machine learning-based approaches have also gained traction, for instance, SVM, Random Tree, KNN, etc. The study[4], [5] proposed supervised learning to detect various cyber-attacks using GOOSE and Sampled Value data. Jay et al. [6] employed unsupervised learning DBSCAN and autoencoder to detect intrusion in GOOSE messages. In [7], a network-level system for efficient detection of GOOSE-based poisoning attacks was proposed to monitor the substation's local area network through port mirroring on switches.

Although the above research has achieved promising results in detecting anomalies and attacks, most of them rely on processed GOOSE datasets, which often strip away the rich semantic and structural information present in the original parsed data (e.g., JSON format). The parsed GOOSE message often contains diverse data types (e.g., timestamps, strings, numeric values, etc.) as shown in Fig. 1 (data in JSON format). For instance, "Activity" is "inform," indicating normal communication. Similarly, values like "SIP3_Abgang3MEAS/LLN0$Meas-000" that deviate from historical ranges may signal attacks or equipment failures. These GOOSE messages are half-structured and have their semantics, which are similar to natural language corpus. However, for such data, traditional machine learning models may require predefined features or feature engineering by experts, neglecting the rich semantic information, thus leading to a significant loss of critical attack indicators. There are many Large Language Models (LLMs) such as ChatGPT [8], Llama [9], Mistral [10], BERT, etc., which are effective in extracting semantic information from text data. Recent studies have leveraged pre-trained and fine-tuned Language Models for anomaly detection in various domains, such as analyzing log data [11] and validating or interpreting financial anomalies [12]. Inspired by these works, we propose to leverage pre-trained LLM to extract semantic and structural features from GOOSE messages, which can capture rich information and

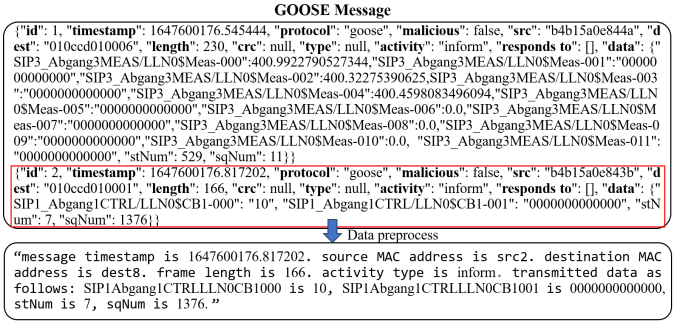


Fig. 1: The illustration of parsed Goose-message.

detect potential attack indicators.

Another issue is data imbalance, where normal samples or some anomaly class dominate the dataset, while others are severely underrepresented. This imbalance increases the risk of overfitting to frequent classes, reducing the model's ability to generalize. Therefore, it is essential to address data imbalance to improve the model's performance in detecting anomalies. In this paper, we target this problem of few-shot anomaly detection. Few-shot learning scenarios, where only a handful of labeled samples per anomaly class are available, further complicate precise anomaly type recognition. Paper [13] proposed a Siamese convolutional neural network to mitigate overfitting and enhance few-shot learning anomaly detection in industrial CPS. Unlike these studies, GOOSE anomaly detection involves both textual and numerical features, making it distinct and more complex.

Moreover, existing anomaly detection techniques are often evaluated only on known anomaly classes seen during training. In practice, models are often required to evaluate samples in an open-world setting: that is, to detect anomalies in previously unseen and novel classes. Despite its significance, cross-category generalization has received limited attention in the literature. Especially in cyber-security area, we may fail to identify the new unknown attacks because we have no patterns or signatures of the attacks to feed into a supervised classifier. Zero-shot learning methods aim to generalize to unseen classes without direct training, as seen in paper [14], which applied unsupervised approach for texture image anomalies. In this work, we target this problem of GOOSE zero-shot anomaly detection, where the model is trained to detect anomalies in the "Insertion attack", "Flooding attack" and "Suppression" classes and expected to generalize to detect anomalies in the previously unseen anomaly classes (e.g, "Replay attack") without further training.

In this paper, we propose a novel framework, GOOSE Attention-based Multi-gate Mixture-of-Experts with a Pre-Trained LLM (GAMMPT), to address the identified challenges. GAMMPT first leverages a pre-trained LLM to extract semantic and structural information from GOOSE messages. It then employs an attention-based Multi-gate Mixture-of-Experts (attMMoE) model to capture anomaly distinctions and model task relationships to identify connections between different anomaly patterns to improve the accuracy in few-shot learning scenarios. For zero-shot, GAMMPT utilizes contrastive learning with clustering to enhance the embedding robustness. Our contributions are as follows:

1) We proposed a novel GAMMPT framework, achieving state-of-the-art performance on real-world GOOSE datasets.
2) We leverage pre-trained LLMs to extract semantic and structural features from semi-structured GOOSE data. To the best of our knowledge, this is the first work applying LLMs for GOOSE anomaly detection.
3) We introduce an attMMoE model, which formulates task relationships and mitigates data imbalance through dynamic expert selection.
4) We develop a contrastive learning strategy with clustering to enable robust zero-shot anomaly detection.

II. Preliminaries and Problem Formulation

1) Data Preprocessing: In this work, we use PowerDuck [15], which contains 30,000 logs from IEC 61850 GOOSE network traffic in a real substation testbed. Unlike many existing synthetic datasets, PowerDuck offers more realistic industrial data, making it a valuable complement for substation security research. The dataset includes 14 anomaly classes and normal traffic. The dataset is imbalanced, with some anomaly classes being significantly underrepresented. For example, 'replay-old-measurements' has only 82 samples, while 'flood-repeat' contains up to 10,278 samples. To prepare the data for GAMMPT, we removed null or irrelevant fields such as 'response to,' 'crc,' and 'id.' Fields with identical values across all records, such as 'transcriber-config,' 'protocol,' and 'type,' were also discarded. Source ('src') and destination ('dest') addresses are string-mapped to 'src(dest)' +'i' for consistency. Malicious labels are converted to integers, {'Normal': 0, 'sup-2': 1, 'insert-distort-meas-up-grad': 2, 'flood-bloat-repeat': 3, 'insert-distort-meas-up-sharp': 4, 'insert-distort-meas-down-sharp': 5, 'sup-1-1-tbv0': 6, 'sup-1-1-tbv2': 7, 'insert-distort-meas-down-grad': 8, 'sup-1': 9, 'sup-1-1-tbv1': 10, 'flood-repeat': 11, 'replay-old-measurements': 12, 'insert-fake-open-w-intermediate': 13, 'insert-fake-open-only-end': 14} (for more information on each anomaly class, see [15]). While in 'data', we replace ':' with 'is' and remove special symbols such as '-', ' ', '/', and '$.' All values are converted to strings, column names are standardized to match the GOOSE protocol's feature name dictionary, for instance: "src": "source MAC address", "dest": "destination MAC address", "length": "frame length", "activity": "activity type". The preprocessed data can be seen in Fig. 1.

2) Problem Formulation: GOOSE anomaly detection task is a multi-class classification problem. The task is formulated as follows: $f : x_i \rightarrow y_i$, where x_i is the i-th GOOSE message, $y_i = \{0, 1, ..., 14\}$ is the anomaly label of x_i.

III. Methodology

This section describes the GAMMPT framework in detail, including the core components: pre-trained LLMs, the attMMoE model, and contrastive learning with clustering for zero-shot learning. The GAMMPT framework is illustrated in

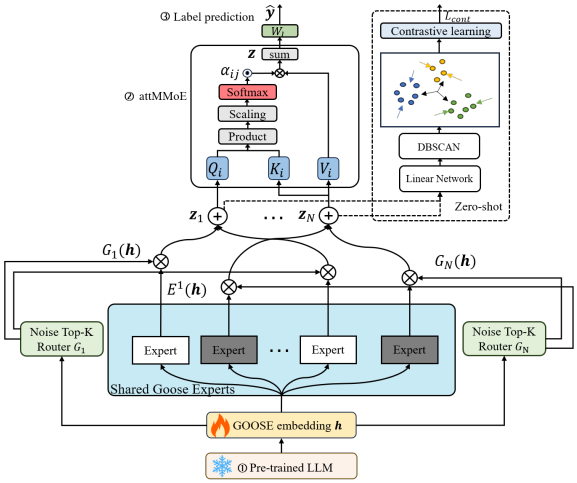


Fig. 2: The GAMMPT framework.

Fig. 2. Each component of the framework is assigned a stage number in the figures for clarity.

1) **Pre-trained LLM (Step 1):** The pre-trained LLM (Distilbert in this work) is used to extract semantic and structural features from the GOOSE messages. After encoding, the LLM output of x_i is represented as h_i.
2) **AttMMoE learning(Step 2):** To improve the accuracy of anomaly type recognition, we decompose the multi-class detection task into multiple binary classification tasks. The shared GOOSE expert and the attention-based gate fusion mechanism are used to capture anomaly distinctions and model task relationships, which can improve the model's interpretation and robustness.
3) **Label prediction (Step 3):** The outputs from attMMoE are passed to the prediction module (constructed by linear networks), to generate predicted labels for GOOSE messages.

A. Pre-trained LLMs

LLMs have shown unprecedented inference ability in many downstream NLP tasks. Thus, we expect them to be effective in extracting semantic and structural features from GOOSE messages. In this work, we leverage DistilBERT [16] as the backbone for our semantic feature extractor due to its balance of efficiency and effectiveness. DistilBERT, a lightweight version of BERT, offers faster training and inference while retaining much of BERT's language understanding capabilities, making it practical for real-world applications. It should be noted that we only utilize the embedding of the pre-trained DistilBERT while keeping the LLM model frozen. The output of the pre-trained DistilBERT is denoted as $\boldsymbol{h} = DistilBERT(\boldsymbol{x}), \boldsymbol{x} = [x_i, \cdots, x_{i+b}]$, where b denotes the batch size.

B. AttMMoE Learning

While the embedding h_i effectively represents general semantic information, they may struggle to distinguish fine-grained anomaly types due to the complex and overlapping feature space in GOOSE messages, especially in highly imbalanced datasets. To address this, we propose the attMMoE model to enhance anomaly type recognition.

First, we decompose the multi-class detection task into multiple binary classification tasks. Since the patterns and features of each anomaly category are usually quite different, it is easier to capture these features by processing them individually. Through multi-task learning, the attMMoE model can simultaneously capture diverse anomaly characteristics, which can improve the model's capacity to handle data imbalance. In this work, we set up 15 tasks for few-shot learning, each of which is a binary classification task. For example, the first task is to distinguish between normal and "sup-2" anomalies, the second task is to distinguish between normal and "insert-distort-meas-up-grad" anomalies, and so on. For the zero-shot learning task, we set up 2 tasks.

The attMMoE model consists of two main components: a shared GOOSE Experts and an attention-based gate fusion mechanism.

1) Shared GOOSE Experts: In our study, we use the Mixture-of-Expert (MoE) model [17] to handle multiple anomaly detection tasks. We share a group of GOOSE experts, and the number of experts is set as M, the number of task is set as T, N represents the total number of samples. Each expert E^j is a neural network designed to capture task-specific anomaly features. The core of the MoE model is a task-specific router that dynamically selects the most relevant experts based on the input features. Specifically, for each task ι, we use a learned Noisy Top-k gating network (G_ι) as a personalized router, which assigns a relevance score to each expert based on the input $\boldsymbol{h}$. The output ($\boldsymbol{z}_\iota$) of the ι-th task MoE is calculated by aggregating the contributions of the top-k selected experts:

$$\boldsymbol{z}_\iota = \sum_{j=1}^{k} G_\iota(\boldsymbol{h}) E^j(\boldsymbol{h}). \tag{1}$$

For G_ι, to encourage exploration during training, we introduce tunable noise into the gating process and then keep the top k values to select the most important k experts. The relevance score for the expert is computed as follows:

$$H_\iota(\boldsymbol{h}) = (\boldsymbol{h} \cdot W_{g_\iota}) + SN(Softplus((\boldsymbol{h} \cdot W_{noise_\iota}))), \tag{2}$$

where W_{g_ι} and W_{noise_ι} are learnable parameters, and the SN function adds Gaussian noise for exploration. We then apply the Top-k selection:

$$Top-k(v,k) = \begin{cases} v_i, & \text{if } v_i \text{ is among the top-}k \text{ values in } v, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The output is activated by a Softmax function and references the weights assigned to each expert. Each expert will then contribute to the final output based on the weights:

$$G_\iota(\boldsymbol{h}) = Softmax(Top-k(H_\iota(\boldsymbol{h}), k)). \tag{4}$$

Remark 1: We only select the top-k relevant experts to improve inference speed and reduce computational costs. The

gating network in a standard MoE setup tends to activate the same few experts, leading to inefficient training. To encourage balanced expert utilization and avoid the collapse of expert selection in standard MoE, we adopt a Noisy Top-k gating strategy. Gaussian noise promotes exploration across experts, while Softplus ensures smooth and positive gating scores. This combination improves training efficiency and expert diversity without compromising task relevance.

2) Attention-based Gate Fusion Mechanism: While traditional MoE utilizes exclusive experts to learn different feature classes, single-task learning models often fail to capture the full complexity of GOOSE-embedded features. To address this limitation, we introduce the attMMoE model, which combines a multi-gate mixture of experts (MMoE) [18] and an attention-based gate fusion mechanism to enable feature sharing across tasks. Specifically, the attention-based gate fusion mechanism aggregates outputs from decoupled anomaly detection tasks, thus enabling our model to share relevant information across tasks. The attention mechanism dynamically fuses these decoupled features, which can help the model determine the correct anomaly type based on task-specific relevance. The attention calculation of the attMMoE is as follows:

$$\boldsymbol{Q}_i = W_q \boldsymbol{z}_i + \boldsymbol{b}_q, \tag{5}$$

$$\boldsymbol{K}_j = W_k \boldsymbol{z}_j + \boldsymbol{b}_k, \tag{6}$$

$$\boldsymbol{V}_j = W_v \boldsymbol{z}_j + \boldsymbol{b}_v, \tag{7}$$

where $\boldsymbol{Q}_i$, $\boldsymbol{K}_j$, $\boldsymbol{V}_j$ are the query, key, and value vectors of the i-th and j-th tasks, respectively. W_q, W_k, W_v are projection matrix for queries Q, keys K and values V, and $\boldsymbol{b}_q$, $\boldsymbol{b}_k$, $\boldsymbol{b}_v$ are the bias terms. The attention score between task i and j is calculated as: $\alpha_{ij} = \frac{\boldsymbol{Q}_i \boldsymbol{K}_j^T}{\sqrt{d_k}}$, where d_k is the dimension of the key vectors. The attention score is then normalized using the Softmax function, $\alpha_{ij} = Softmax(\alpha_{ij})$. The output of the attMMoE is calculated as: $\boldsymbol{z} = \sum_{j=1}^{T} \alpha_{ij} V_j$. Then, the output $\boldsymbol{z}$ is input into the prediction module to predict the anomaly label. $\hat{\boldsymbol{y}} = W_l \boldsymbol{z} + b_l$, where W_l and b_l are the learnable weight and bias of the linear layer, respectively. The cross-entropy loss of target label y and predicted label $\hat{y}$ is calculated as:

$$L_{cross-entropy} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i). \tag{8}$$

C. Contrastive learning with clustering

To enhance the model's practicality, we address a zero-shot learning scenario where GAMMPT learns prior knowledge about both normal and abnormal behaviors to detect unseen anomalies. To achieve this, we propose integrating contrastive learning and clustering to improve embedding robustness (as seen in the 'zero-shot' part of Fig. 2). During the learning stage, DBSCAN clusters GOOSE embeddings generated by attMMoE, and the clustering results are used to assign pseudo-labels in an unsupervised manner. The number of clusters is C. Embeddings from the same cluster are denoted as positive pairs $(\boldsymbol{z}_i^{'}, \boldsymbol{z}_j^{'})$, while embeddings from different clusters are negative pairs $(\boldsymbol{z}_i^{'}, \tilde{\boldsymbol{z}}_j^{'})$. These pseudo-labels are then used for contrastive learning, which maximizes the agreement between embeddings from the same cluster while separating dissimilar ones. This approach strengthens representation learning by aligning embeddings of related anomalies while distinguishing unrelated ones. As a result, the model achieves better generalization ability for detecting unseen anomaly classes. The contrastive loss is calculated as:

$$L_{cont} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{exp(sim(\boldsymbol{z}_i^{'}, \boldsymbol{z}_j^{'})/\tau)}{\sum_{j=1}^{N} exp(sim(\boldsymbol{z}_i^{'}, \tilde{\boldsymbol{z}}_j^{'})/\tau)}. \tag{9}$$

The contrastive loss is computed to optimize this embedding space by pushing positive pairs closer and negative pairs farther apart, enhancing the model's robustness in zero-shot anomaly detection.

IV. Experimental Results

In this section, case studies are conducted on the PowerDuck dataset [15]. Our GAMMPT is first compared against several baselines in few-shot learning and zero-shot learning. Then, an ablation study analysis is conducted, accompanied by visualization analyses.

A. Baseline

We compare the proposed GAMMPT with several baselines, including Design tree, SVM, KNN, MLP, LSTM, and fine-tune DistilBERT in a few-shot learning scenario. For a zero-shot learning scenario, we also compare the GAMMPT with Llama and Mistral, which are state-of-the-art models for zero-shot learning. All baselines are trained with the same GOOSE embeddings as GAMMPT, which is the output from DistilBERT.

B. Experimental Setups

In the experiments, the batch size for all models is set to 16. The learning rate for other baselines is set to 1e-4, GAMMPT uses 1e-5 as the learning rate, while the fine-tuning DistilBERT, LLaMA, and Mistral use a learning rate of 5e-5. The dropout rate is fixed at 0.3 for all models. In the attMMoE model, the number of experts M is set to 16, with the top-k selection set to 4. For the contrastive learning with clustering module, the number of clusters C in DBSCAN is set to 5. The temperature parameter τ in the contrastive loss is set to 0.1. We evaluate few-shot learning using a 10-shot setting. For zero-shot learning, the model is trained on 5 anomaly classes (e.g., labels 0-5) and tested on 10 unseen anomaly classes (labels 6-14). The experiments ran for 200 communication rounds, selecting the model with the minimum global validation loss as the final model. Hyperparameter settings, including learning rates {1e-5, 5e-5, 1e-4}, batch sizes {16, 32, 64}, and dropout rates {0, 0.3, 0.5, 0.7}, are optimized using GridSearch.

All comparative experiments are implemented using Python 3.10 on an Nvidia A100 GPU. The evaluation metrics used are weighted precision, weighted recall, and weighted F1-score, which accounts for class imbalance by using the sample size of each category as a weight.

TABLE I: Few-shot learning anomaly detection performance comparison.

Model	Precision	Recall	F1-score
DistilBERT(frozen)+Design tree	0.7475	0.5931	0.5155
DistilBERT(frozen)+KNN	0.9768	0.9768	0.9704
DistilBERT(frozen)+SVM	**0.9866**	0.9814	0.9817
DistilBERT(frozen)+MLP	0.9652	0.9455	0.9383
DistilBERT(frozen)+LSTM	0.7357	0.0586	0.1046
DistilBERT (fine tune)	0.9461	0.9366	0.9355
GAMMPT	0.9852	**0.9819**	**0.9824**

C. Few-shot Learning Anomaly Detection Performance Comparison

In the training dataset, normal data is split into training (70%), validation (15%), and testing (15%) sets. Each anomaly class has 10 samples reserved for training, with the remainder evenly split between validation and test sets. The few-shot learning results are summarized in Table I. Decision Tree performs poorly due to its limited capacity to handle complex feature representations from frozen embeddings. KNN and SVM outperform fine-tuned DistilBERT, which shows that combining pre-trained embeddings with appropriate classifiers can enhance anomaly detection while fine-tuning alone is insufficient. Among neural network-based models, MLP surpasses LSTM. LSTM likely overfits due to limited training samples. This indicates that overly complex models may struggle in few-shot scenarios. GAMMPT achieves the best performance, with a weighted precision of 98.52%, recall of 98.19%, and F1-score of 98.24%, outperforming most of the other baselines. Its superior results demonstrate the effectiveness of the MMoE framework for shared learning across tasks and attention-based gate fusion (attMMoE) for aggregating task-specific features. These mechanisms enable GAMMPT to capture feature relationships and improve task-specific representations, making it highly effective for anomaly detection in few-shot learning scenarios.

D. Zero-shot Learning Anomaly Detection Performance Comparison

In the zero-shot learning setting, data labeled 0-5 is designated as known, while data labeled 6-14 represents unknown anomalies. The known data is divided into a testing set (70%), with the remainder evenly split into training and validation sets. The final testing set includes unseen known samples (0-5) and all unknown anomaly samples (6-14). The results, summarized in Table II, show that most baselines perform poorly in recall and F1-score due to the presence of unseen anomaly classes. However, Llama and Mistral, state-of-the-art LLMs, achieve relatively high precision, recall, and F1 scores. This indicates that LLMs possess inherent inference capabilities gained from pre-training on large-scale datasets, which enable them to generalize better and recognize feature patterns even in unfamiliar anomaly classes. GAMMPT achieves the best performance with weighted precision (95.58%), recall (92.27%), and F1-score (93.18%), outperforming all the baseline models. Compared to the worst result, GAMMT improves in precision, recall, and F1-score 30%, 59%, and 86%, respectively. The

TABLE II: Zero-shot learning anomaly detection performance comparison.

Model	Precision	Recall	F1-score
DistilBERT(frozen)+Design tree	0.7356	0.5811	0.5005
DistilBERT(frozen)+KNN	0.7692	0.5828	0.5034
DistilBERT(frozen)+SVM	0.8681	0.8424	0.8149
DistilBERT(frozen)+MLP	0.8317	0.7803	0.7375
DistilBERT(frozen)+OneClassSVM	0.9363	0.1063	0.1703
DistilBERT (fine tune)	0.917	0.4893	0.5732
Llama (Llama-3-8B)	0.9038	0.8304	0.8561
Mistral (Mistral-7B)	0.861	0.9018	0.8586
GAMMPT	**0.9558**	**0.9227**	**0.9318**

TABLE III: Ablation study of the attMMoE model.

Model	Precision	Recall	F1-score
MoE(FS)	0.9776	0.9728	0.9725
MMoE(FS)	0.9817	0.9784	0.9793
GAMMPT(FS)	**0.9852**	**0.9819**	**0.9824**
MoE(ZS)	0.922	0.646	0.7168
MMoE(ZS)	0.9286	0.7511	0.8008
GAMMPT(ZS)	**0.9558**	**0.9227**	**0.9318**

superior performance of GAMMPT can be attributed to its use of contrastive learning combined with clustering. By leveraging feature similarities between known and unknown anomaly classes, the model effectively clusters similar representations. This can strengthen the robustness of the learned embeddings, and enhance the model's generalization ability in detecting previously unseen anomaly classes.

E. Ablation Study

1) AttMMoE model ablation experiments: To evaluate the effectiveness of the attMMoE model, we conduct ablation studies by comparing GAMMPT with the MoE and MMoE framework in both few-shot learning (FS) and zero-shot (ZS) learning. The results are summarized in Table III. Results show that GAMMPT with MMoE outperforms MoE in both few-shot and zero-shot scenarios, demonstrating that modeling task relationships can enhance the ability to capture anomaly distinctions. The attMMoE model further improves performance by aggregating task-specific features and improving task-specific representations. Compared with the worst result in zero-shot learning, GAMMPT achieves improvements of 3.7%, 42.8%, and 30%, respectively. The results reveal that the attMMoE model is effective in enhancing anomaly detection in few-shot and zero-shot learning scenarios.

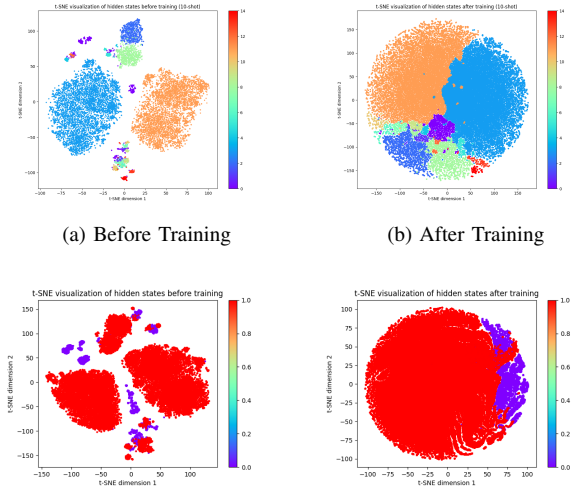
2) Contrastive learning with clustering ablation experiments: To evaluate the effectiveness of contrastive learning with clustering, we conduct an ablation study by comparing GAMMPT with and without contrastive learning in zero-shot learning. As shown in Table III, GAMMPT with contrastive learning outperforms GAMMPT without contrastive learning in precision, recall, and F1-score, improved by 3%, 22.7%, and 16.4%, respectively. This suggests that contrastive learning with clustering can enhance embedding robustness and improve the model's generalization ability in zero-shot anomaly detection.

F. Visualization Analysis

We examined embedding learning from attMMoE by using t-SNE visualization (Fig. 3). Before training, the embeddings

TABLE IV: Ablation study of contrastive learning with clustering.

Model	Precision	Recall	F1-score
GAMMPT(ZS)	0.9283	0.7522	0.8016
GAMMPT(ZS)+Contrastive	**0.9558**	**0.9227**	**0.9318**



(a) Before Training (b) After Training

(c) Before Training (Zero-shot) (d) After Training (Zero-shot)

Fig. 3: Visualization of embedding space using t-SNE.

show partial clustering, but there remain significant blank indicating incomplete representation learning. Anomalies of the same type, such as those in purple, are dispersed across the embedding space rather than being concentrated in a distinct region, reflecting limited latent space regularization. After training, the embedding space becomes more organized and evenly distributed, with distinct boundaries between different anomaly types. This improvement highlights attMMoE's ability to enforce latent space regularization, creating a smooth and continuous embedding space.

Similarly, in the zero-shot learning setting, before training, the embedding space has some overlap between different anomaly classes. attMMoE with contrastive learning further improves the embedding space's structure. Similar embeddings are grouped while dissimilar embeddings are pushed apart, demonstrating improved generalization through contrastive learning. Overall, the combination of attMMoE and contrastive learning leads to efficient feature extraction, enabling the model to learn meaningful, task-specific embeddings while maintaining generalization capabilities.

V. Conclusion

This paper presents GAMMPT, a novel framework for detecting anomalies in GOOSE-based substation communication, designed to enhance the security and resilience of power system operations. By integrating a Pre-Trained Large Language Model with attMMoE, contrastive learning and clustering, GAMMPT effectively identifies diverse cyber anomalies, even under limited labeled data conditions. The approach improves the robustness and generalization capability of anomaly detection models, making it well-suited for deployment in real-world substations. In future work, we aim to develop an LLM-driven knowledge base of communication anomaly patterns to support proactive grid monitoring and situational awareness. Additionally, we plan to further investigate the theoretical foundations of the attMMoE architecture to enhance model interpretability.

References

[1] G. Elbez, K. Nahrstedt, and V. Hagenmeyer, "Early attack detection for securing goose network traffic," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 899–910, 2023.

[2] G. Elbez, H. B. Keller, and V. Hagenmeyer, "A new classification of attacks against the cyber-physical security of smart grids," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 2018, pp. 1–6.

[3] S. M. Farooq, S. S. Hussain, and T. S. Ustun, "Performance evaluation and analysis of iec 62351-6 probabilistic signature scheme for securing goose messages," *IEEE Access*, vol. 7, pp. 32 343–32 351, 2019.

[4] S. Bhattacharya, N. Saqib, and M. Govindarasu, "Ml-based anomaly detection system for iec 61850 communication in substations," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2024, pp. 1–5.

[5] K. Park, M. Girdhar, J. Hong, W. Su, A. Herath, and C.-C. Liu, "Machine learning based cyber system restoration for iec 61850 based digital substations," *arXiv preprint arXiv:2411.07419*, 2024.

[6] D. Jay, H. Goyel, U. Manickam, and G. Khare, "Unsupervised learning based intrusion detection for goose messages in digital substation," in *2022 22nd National Power Systems Conference (NPSC)*. IEEE, 2022, pp. 242–247.

[7] A. Bohara, J. Ros-Giralt, G. Elbez, A. Valdes, K. Nahrstedt, and W. H. Sanders, "Ed4gap: Efficient detection for goose-based poisoning attacks on iec 61850 substations," in *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2020, pp. 1–7.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[9] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[11] C. Almodovar, F. Sabrina, S. Karimi, and S. Azad, "Logfit: Log anomaly detection using fine-tuned language models," *IEEE Transactions on Network and Service Management*, 2024.

[12] T. Park, "Enhancing anomaly detection in financial markets with an llm-based multi-agent framework," *arXiv preprint arXiv:2403.19735*, 2024.

[13] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, "Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5790–5798, 2021.

[14] Y. Li, A. Goodge, F. Liu, and C.-S. Foo, "Promptad: Zero-shot anomaly detection using text prompts," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1093–1102.

[15] S. Zemanek, I. Hacker, K. Wolsing, E. Wagner, M. Henze, and M. Serror, "Powerduck: a goose data set of cyberattacks in substations," in *Proceedings of the 15th Workshop on Cyber Security Experimentation and Test*, 2022, pp. 49–53.

[16] V. Sanh, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[17] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[18] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1930–1939.