A Mixture-of-Agents Framework for EV Battery Diagnostics: Semantic Clustering and Prompt Engineering for Automated Reporting

Junhyung Moon[12*], **Seunggwan Hong**[12*], **Eunkyeong Lee**[12*], **Yunho Lim**[4], **Wanjin Park**[1] and **Hyunseung Choo**[23†]

[1]KT, Seoul, S.Korea
[2]Dept. of AI Systems Engineering
[3]Dept. of Electrical and Computer Engineering
[4]Dept. of Smart Factory Convergence
Sunkyunkwan University, Suwon, S.Korea
{junhyung.moon,seunggwan.hong,ek.lee,wjpark25}@kt.com,
{dbsgh0671,choo}@skku.edu
[*] Equal contributions
[†] Corresponding author

Abstract

Large language models (LLMs) have shown considerable promise for interpreting structured data, yet their use on electric-vehicle (EV) time-series remains limited. We introduce a framework that fuses semantic-aware clustering with prompt engineering to produce diagnostic reports from high-dimensional EV battery logs. Central to our approach is a Mixture-of-Agents (MoA) architecture in which several LLM-driven agents cluster the data from complementary perspectives before their outputs are unified into semantically coherent groups. These clusters then drive a few-shot prompting strategy for report generation. We evaluate three prompting variants that supply progressively richer context, using the LLM-as-Judge protocol. Experiments show that MoA yields higher silhouette scores than K-means, and that prompts enriched with same-cluster samples plus inferred cluster summaries deliver the most informative reports. The results highlight how combining semantic clustering with careful prompt design enhances both interpretability and quality of LLM outputs. This work provides a foundation for automated reporting in real-world EV diagnostics.

1 Introduction

As the adoption of Electric Vehicles (EVs) accelerates, the volume and complexity of time-series data generated during vehicle operation, such as driving logs and battery state, continues to increase. However, this data is inherently high-dimensional and non-linear, and often reflects overlapping influences from diverse driving conditions and environmental factors. These characteristics pose challenges for traditional statistical methods or rule based reporting systems, which often fail to capture complex patterns and anomalies effectively [Li *et al.*, 2019; Steinstraeter *et al.*, 2020]. To address these limitations, we propose a novel pipeline that clusters EV time-series logs and generates diagnostic reports for each cluster using a large language model (LLM). The system first clusters preprocessed time-series data, then produces situation-specific reports for each cluster through LLM-based few-shot prompting.

Traditional clustering methods such as K-means suffer from instability and poor separation due to their sensitivity to initial conditions and fixed partition criteria. To overcome these limitations, we adopt a Mixture of Agents (MoA) architecture [Wang *et al.*, 2024], where multiple LLM based agents independently generate clustering proposals based on different feature subsets or prompting perspectives, and a final aggregation determines the outcome. This structure ensures robust and consistent clustering, even across repeated runs. Clustering is performed using key features relevant to battery behavior, such as average speed and temperature rise (ΔT), and representative samples from each cluster are used as few-shot exemplars in prompts, allowing the LLM to generate high-quality reports without additional training. Such prompting strategies have been proven effective in tasks such as industrial summarization, fault diagnosis, and process monitoring [Chen *et al.*, 2025; Ning *et al.*, 2023; Pu *et al.*, 2024].

To evaluate report quality, we adopt a win/tie/lose comparative judgment scheme inspired by MT-Bench [Zheng *et al.*, 2023], where two reports are presented side-by-side and assessed by an LLM judge. Experimental results show that our MoA based clustering achieved a 43.5% improvement in Silhouette Score over K-means, and our most comprehensive prompting strategy incorporating same cluster samples and LLM-derived cluster descriptions attained a win rate of 67.22% over baseline.

Our contributions are threefold:

- A pipeline is proposed that clusters EV time-series data

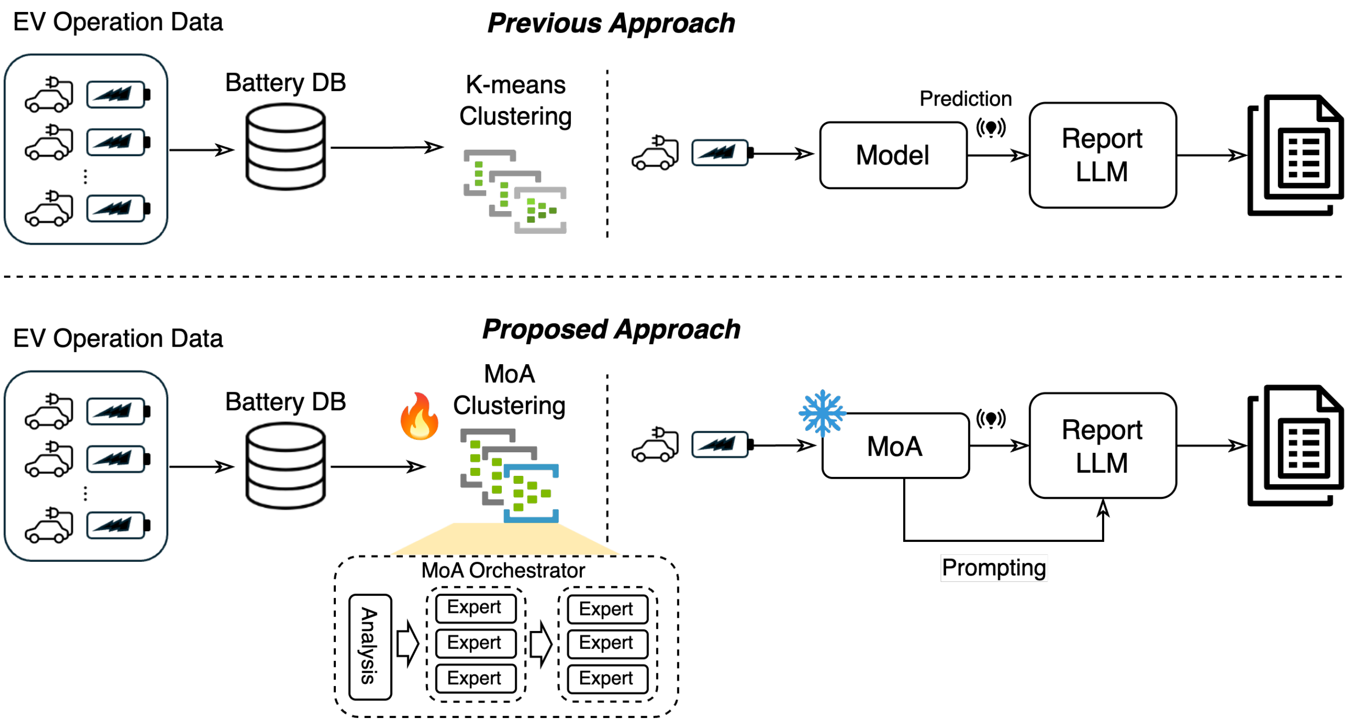


Figure 1: Proposed Method: Baseline K-means Pipeline vs. Proposed MoA-Driven Clustering and Prompting Framework for EV Battery Report Generation.

using key battery-related features and produces cluster level diagnostic reports through few-shot prompting with a LLM.

- To improve clustering quality, a MoA is introduced, wherein multiple agents generate clustering candidates from different perspectives. Its effectiveness over K-means is validated through comparative evaluation.
- We assess report quality using a win, tie or lose judgment framework and show that our prompting method consistently yields higher-quality outputs.

2 Related Works

2.1 Time-Series Data Analysis

Time-series data inherently possesses multi dimensional attributes. Among unsupervised analysis methods for such data, clustering is one of the most commonly used approaches, with the K-means algorithm being the most prevalent due to its simplicity and computational efficiency. It is widely adopted in various time-series data analysis tasks. However, K-means has inherent limitations: it requires the number of clusters to be predefined, is sensitive to initial centroid placement, and relies on distance-based partitioning, which makes it inadequate for capturing complex non-linear structures or interactions among high-dimensional features. These limitations are particularly pronounced in EV log data, where heterogeneous time-series patterns arise from diverse driving conditions and user behaviors.

Tayarani et al. pointed out that traditional clustering methods such as K-means struggle to represent the complexity of EV charging behavior [Tayarani *et al.*, 2023]. Similarly, Ke and Wang demonstrated that segmenting and processing time-series data from multiple perspectives, rather than a single criterion, can significantly enhance prediction accuracy and flexibility [Ke and Wang, 2024]. This perspective is structurally aligned with the philosophy behind the MoA approach proposed in our work, which leverages LLMs for multi-perspective clustering.

2.2 Time-Series Data with LLM

LLMs have recently been applied with increasing frequency to interpret or summarize time-series data in natural language. In previous studies showed that by converting numerical time-series into character token sequences, LLMs can perform even zero-shot forecasting tasks [Gruver *et al.*, 2023]. Hegselmann et al. proposed a method that serializes tabular data into textual form, enabling LLMs to perform classification and summarization tasks on otherwise structured input [Hegselmann *et al.*, 2023]. These studies suggest that LLMs are capable of understanding non-standard input formats and generating domain-specific responses.

In such tasks, performance is heavily influenced by how structured data is transformed and fed into LLMs. Recently, semantically relevant example selection for few-shot prompting has emerged as a dominant strategy [Achiam *et al.*, 2023; Gruver *et al.*, 2023]. For instance, Mohan et al. demonstrated significant improvements in named entity recognition (NER) within the medical domain by selecting few-shot exemplars via K-means clustering [Mohan *et al.*, 2024]. Inspired by these approaches, our study extends this few-shot prompting methodology to the EV time-series domain.

2.3 LLM as a Judge

When evaluating outputs generated by language models, traditional quantitative metrics (e.g., BLEU, ROUGE) are often insufficient to capture expressive diversity and multi-dimensional quality. As a result, comparative evaluation methods especially those simulating human level judgment have gained attention. A representative example is MT-Bench, proposed by Zheng et al., which employs a win/tie/lose evaluation scheme by presenting two responses side-by-side and asking either an LLM or a human to choose the better one, or declare a tie [Zheng *et al.*, 2023].

Zheng et al. showed that GPT-4, when used as a judge in such evaluations, achieved an agreement rate of up to 85% with human annotators, thereby demonstrating the reliability of LLMs as qualitative evaluators. In our study, we adopt this evaluation framework to compare the quality of reports generated from the same EV time-series input, using either MoA based or K-means based clustering. The relative quality of these reports is quantitatively assessed via pairwise comparisons under the win/tie/lose scheme.

3 Methodology

We hypothesize that entropy loss, which often occurs when clustering high-dimensional time-series data using methods like K-means, can be mitigated by leveraging the semantic capabilities of LLMs. To this end, we propose a method that improves the transformation of structured time-series data into unstructured diagnostic reports, validated through a report generation task based on real-world EV battery management. Figure 1 provides an overview of both the baseline and our proposed pipeline. In the baseline (top), EV battery time-series data are clustered using K-means, followed by a prediction model that produces input for the LLM to generate the report.

Each stage operates independently, and clustering results are not directly used in the report generation process. In contrast, our method (bottom) adopts a MoA, where multiple agents with different criteria perform clustering in parallel. Their outputs are aggregated to produce a more stable and semantically meaningful cluster structure, marked with a fire icon (activated). From these clusters, a few-shot prompting context is constructed, while the LLM itself remains fixed and generates the report based on the given prompt, as indicated by a snow icon (frozen). We used the 'gpt-4o' and 'claude-4' LLM for both report generation and evaluation. This architecture enables tighter integration between clustering and prompting while ensuring stability in the generation stage.

3.1 Clustering

Traditional distance based clustering methods, such as K-means, face limitations when applied to high-dimensional time-series data due to fixed similarity metrics and sensitivity to initialization. Even techniques like dynamic time warping (DTW) offer limited ability to capture domain specific semantics [Dhillon *et al.*, 2004]. To address these shortcomings, we adopt a MoA framework, which leverages the semantic flexibility of LLMs to enable clustering from multiple perspectives [Wang *et al.*, 2025].

The MoA framework comprises an Analysis Agent, multiple Worker Agents, and an Orchestrator that convert raw battery logs and a user query into interpretable clusters of SoC trajectories. MoA operates through three cooperative modules. First, the LLM-driven Analysis Agent performs automated feature attribution, selecting the variables (e.g., voltage–current profiles, temperature, and internal resistance) that exert the greatest influence on SoC and embedding them in domain-specific prompt templates. Next, three Worker Agents execute a two-layer clustering cascade. In Layer 1, each agent explores the feature space using distinct distance metrics, random initialisations, and hyperparameters, yielding candidate partitions. In Layer 2, the Orchestrator evaluates internal and external validity indices, refines hyperparameters or cluster counts, and instructs the Workers to re-cluster. Finally, the Orchestrator aggregates the revised partitions into a consensus assignment that reconciles statistical structure with insights from battery science. This pipeline produces clusters that both enhance downstream SoC-prediction models and provide transparent explanations of the battery attributes that define each group.

3.2 Report Generation

To generate unstructured diagnostic reports from structured time-series data, we propose a prompting strategy tailored for LLMs. Our method utilizes clustering results to incorporate domain specific information into the prompt, enabling automatic generation of EV battery management reports intended for practitioners. Unlike traditional approaches that directly convert structured data into text, our strategy actively leverages the underlying cluster structure to improve both the informativeness and consistency of generated reports. We design prompts with varying levels of cluster information and empirically compare their effectiveness. Each prompt is designed to help the LLM reason about how features such as average speed, temperature rise, and HVAC usage affect SoC consumption. The goal is to incrementally enhance the depth, coherence, and factual relevance of the generated reports. The full prompt texts are provided in Table 1.

- Basic: A report is generated using only the structured input converted into JSON format.
- Ours (+cluster sample): In addition to the input, samples from the same cluster are included.
- Ours (+cluster sample & info): Same cluster samples are supplemented with LLM inferred cluster level descriptions.

4 Experimental Results

4.1 Dataset

We employ real-world driving data from a publicly available BMW i3 dataset [Steinstraeter *et al.*, 2020], comprising 72 multivariate time-series sessions that capture battery and thermal behaviour under diverse external and internal

Type	Prompt
Basic	As a battery expert, you are a helpful assistant who can provide a detailed and complete battery characterisation & management report based on the EV characteristics data for a given drive. You should consider how each feature affects SoC consumption, and make sure to account for key features (average speed, temperature rise, air conditioning/heater use, etc.). This report should be helpful to practitioners of electric vehicle battery management. Driving data: {test_datapoint}
Basic+cluster_sample	As a battery expert, you are a helpful assistant who can provide a detailed and complete battery characterisation & management report based on the EV characteristics data for a given drive. You should consider how each feature affects SoC consumption, and make sure to account for key features (average speed, temperature rise, air conditioning/heater use, etc.). This report should be helpful to practitioners of electric vehicle battery management. Driving data: {test_datapoint} Data from the same cluster as the driving data: {same_cluster_datapoint_sample}
Ours(Basic+cluster_sample+cluster_info)	As a battery expert, you are a helpful assistant who can provide a detailed and complete battery characterisation & management report based on the EV characteristics data for a given drive. You should consider how each feature affects SoC consumption, and make sure to account for key features (average speed, temperature rise, air conditioning/heater use, etc.). This report should be helpful to practitioners of electric vehicle battery management. Driving data: {test_datapoint} Data from the same cluster as the driving data: {same_cluster_datapoint_sample} Cluster properties: {cluster_info_from_LLM}

Table 1: Prompts used in the report generation pipeline.

Feature	Unit	Description
Battery Temperature	°C	Internal battery pack temperature (sensor reading)
State of Charge (SoC)	%	Percentage of remaining battery capacity
SoC Difference	%	SoC change between session start and end
Ambient Temperature	°C	External air temperature (sensor reading)
Target Cabin Temperature	°C	Desired interior temperature set by driver
Distance	km	Total driving distance during session
Duration	min	Total driving time during session

Table 2: Summary of dataset features used for clustering and analysis.

Trip	Date	Route/Area	Weather	Batt Temp (Start)	Batt Temp (End)	SoC (Start)	SoC (End)	SoC Diff	Ambient Temp	Cabin Temp	Distance [km]	Duration [min]	Fan
TripA01	2019-06-25_13-21-14	Munich East	sunny	21.0	22.0	0.863	0.803	0.060	25.5	23.0	7.43	16.82	Auto, L1
TripA02	2019-06-25_14-05-31	Munich East	sunny	23.0	26.0	0.803	0.673	0.130	32.0	23.0	23.51	23.55	Auto, L1
… (more rows follow)													

Table 3: Trip data example.

conditions. Key features such as battery temperature, state of charge (SoC), SoC variation, ambient and cabin temperatures, trip distance, and trip duration, characterise battery performance and energy consumption. They serve as inputs for both clustering and report generation, as summarised in Table 2. dataset example is Table 3.

4.2 Clustering Accuracy

The proposed MoA based clustering model is compared with the conventional K-means algorithm on battery data using key feature subsets. Cluster quality is evaluated via the silhouette score, which measures intra-cluster cohesion (how closely points group together) and inter-cluster separation (how far

Cluster	Information
A	Short range, low SoC consumption - Features: Short trips, low SoC consumption, low battery temperature variation - This cluster consists of trips with short distances and low SoC consumption. - It is mainly composed of vehicles used for short trips in the city centre or for commuting.
B	Medium range, medium SoC consumption - Features: Medium mileage, medium SoC consumption, medium battery temperature variation - This cluster consists of trips with medium range and medium SoC consumption. - The battery temperature variation is also moderate. It mainly includes trips outside or near city centres.
C	Long range, high SoC consumption - Features: Long range, high SoC consumption, large battery temperature fluctuations - This cluster consists of trips with long distances travelled and high SoC consumption. - The battery temperature variation is also large. It mainly includes long distance driving or highway driving.

Table 4: Cluster information as delimited by MoA.

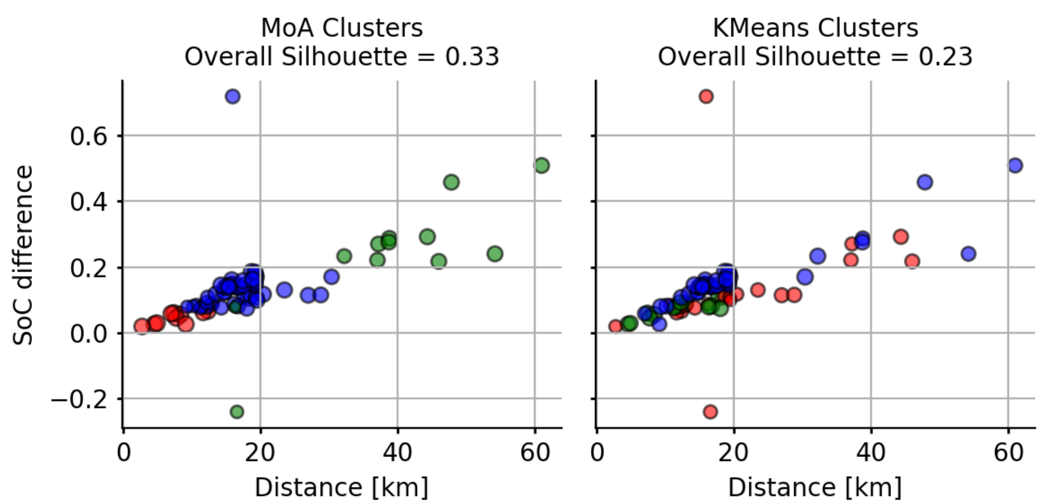


Figure 2: Baseline K-means Pipeline vs. Proposed MoA-Driven Clustering and Prompting Framework for EV Battery Report Generation.

Prompts
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question below. You should choose the assistant that better follows the user's instructions and answers the question. Your evaluation should consider factors such as helpfulness, relevance, accuracy, depth, creativity, and level of detail. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the response order does not influence your decision. Do not let the length of the responses affect your evaluation. Do not favor specific assistant names. Be as objective as possible. After your explanation, provide your final verdict in this format: "A" if assistant A is better, "B" if assistant B is better, and "C" for a tie. Report A: {Report_A} \n\n Report B: {Report_B}

Table 5: Prompts used in the judgement.

Judge LLM	Model	Comparison Model	Win	Tie	Lose
GPT-4o(20240718)	**Ours(+cluster sample & info)**	Basic	**67.22**	18.89	13.89
	Ours(+cluster sample)	Basic	62.22	19.44	18.33
	Ours(+cluster sample)	**Ours(+cluster sample & info)**	31.11	32.22	**36.67**
Claude-sonnet-4(20250514)	**Ours(+cluster sample & info)**	Basic	**77.22**	1.1	21.67
	Ours(+cluster sample)	Basic	71.67	2.2	26.11
	Ours(+cluster sample)	**Ours(+cluster sample & info)**	16.67	36.67	**47.22**

Table 6: Pairwise MT-Bench Evaluation of Prompting Variants.

they are from neighboring clusters). This provides a quantitative measure of how well each point fits within its cluster and how distinct it is from others.

To compute the score, the average intra-cluster distance $a(i)$ is calculated for each data point i, where j denotes another sample, $d(i,j)$ is the Euclidean distance between points i and j, C is the cluster to which i belongs, and C' is a cluster that does not contain i. The separation $b(i)$ is computed as the minimum average distance to points in the nearest neighboring cluster. The silhouette score $s(i)$ is then defined as:

$$a(i) = \frac{1}{|C|-1} \sum_{\substack{j \in C \\ j \neq i}} d(i,j) \quad (1)$$

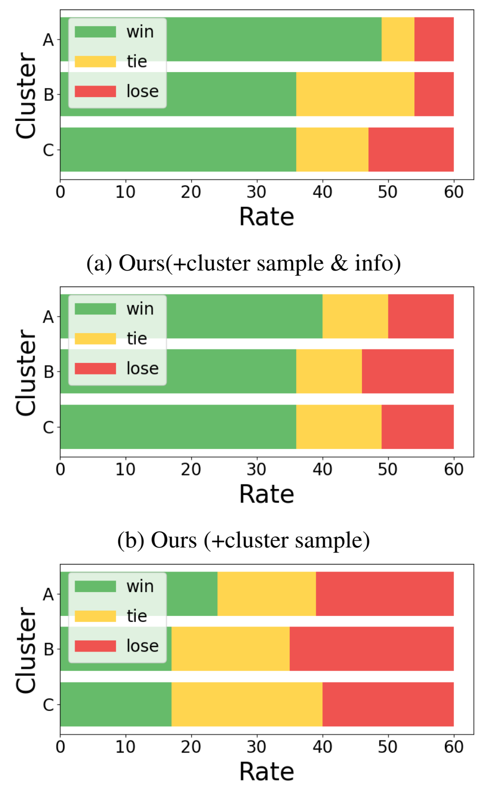
$$b(i) = \min_{C' \neq C} \frac{1}{|C'|} \sum_{j \neq C'} d(i,j) \quad (2)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

Figure 2 presents the clustering results based on real-world data, comparing the proposed model with the conventional K-means algorithm. When evaluated using the silhouette score, the MoA based clustering outperformed K-means by approximately 43.5%. This result demonstrates the effectiveness of the proposed MoA clustering method over the traditional K-means approach. Detailed characteristics of each MoA cluster are provided in Table 4.

4.3 Report Judgement

To compare the quality of reports produced by different prompting strategies, we adopt the MT-Bench framework with the LLM-as-Judge protocol. Reports generated under each of the three strategies are paired and scored by the LLM as win, tie, or lose. Because cluster statistics differ, each cluster (A, B, C) is evaluated independently with 60 held-out samples, and response order is randomized to remove positional bias. The results appear in Figure 3. Figure 3 (a) shows that the prompt containing both same cluster samples and LLM inferred cluster summaries attains the highest win rate across all clusters, indicating that richer context yields more informative and accurate reports. For comparison, Figure 3 (b) reports outcomes when only same cluster samples are supplied. We also directly contrast our two proposed prompting variants in Figure 3 (c). Although the overall differences are modest, the variant that additionally provides LLM generated cluster summaries consistently secures a slight edge,



(a) Ours(+cluster sample & info)

(b) Ours (+cluster sample)

(c) Ours (+cluster sample) vs Ours(+cluster sample & info)

Figure 3: Win/Tie/Lose Outcomes by Prompt Variant with GPT-4o.

suggesting that including every available piece of contextual information is beneficial for report quality. Full numerical results are listed in Table 6, and the exact evaluation prompts appear in Table 5. We used gpt-4o and cluade-4 as judge llm. Boldface numbers denote the best performance within each comparison.

5 Conclusion

This paper proposes a novel framework that clusters high-dimensional EV time-series data and generates diagnostic reports using LLMs. The method introduces a MoA architecture, where multiple agents perform clustering from diverse perspectives based on LLM reasoning, and their outputs are integrated into semantically coherent clusters. These clusters are then used to construct prompts that gradually expand contextual information during report generation.

The effectiveness of the proposed method is empirically validated in two aspects. First, MoA based clustering outperforms the traditional K-means algorithm in terms of silhouette score, demonstrating improved cohesion and separation. Second, the MT-Bench evaluation based on the LLM-as-Judge protocol confirms that prompts containing both same cluster samples and LLM inferred cluster summaries result in the highest quality reports. These findings highlight that integrating semantic-aware clustering with prompt engineering enhances the interpretability and practicality of LLM outputs in structured time-series domains.

For future work, we plan to enhance the MoA framework by explicitly incorporating temporal dependencies, aiming to improve clustering performance on long or complex time-series data. We also intend to broaden comparative studies across various clustering algorithms and evaluation metrics to further verify the generalizability and robustness of the proposed approach. In addition, we aim to develop an interactive system that continuously refines clustering criteria and prompt design through feedback from domain experts, enabling ongoing performance improvements and deeper insight generation in real-world EV applications. Lastly, we plan to deploy this system as a practical EV battery diagnostic service and extend it into an integrated reporting solution suitable for industrial environments.

Acknowledgments

This work was partly supported by the Korea government (MSIT), IITP, Korea, under the ICT Creative Consilience program (IITP-2025-RS-2020-II201821, 40%) and Development of 6G Network Integrated Intelligence Plane Technologies (RS-2024-00392332, 30%); and by the National Research Foundation of Korea(NRF) under grant (RS-2024-00343255, 30%).

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Chen *et al.*, 2025] Jiao Chen, Ruyi Huang, Zuohong Lv, Jianhua Tang, and Weihua Li. Faultgpt: Industrial fault diagnosis question answering system by vision language models. *arXiv preprint arXiv:2502.15481*, 2025.

[Dhillon *et al.*, 2004] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 551–556, New York, NY, USA, 2004. Association for Computing Machinery.

[Gruver *et al.*, 2023] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635, 2023.

[Hegselmann *et al.*, 2023] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

[Ke and Wang, 2024] Fucai Ke and Hao Wang. Divide-conquer transformer learning for predicting electric vehicle charging events using smart meter data. *arXiv preprint arXiv:2403.13246*, 2024.

[Li *et al.*, 2019] Xuefang Li, Qiang Zhang, Zhanglin Peng, Anning Wang, and Wanying Wang. A data-driven two-level clustering model for driving pattern analysis of electric vehicles and a case study. *Journal of cleaner production*, 206:827–837, 2019.

[Mohan *et al.*, 2024] Meethu Mohan, Sneha Shaji Punnan, and Jeena Kleenankandy. Improving few-shot prompting using cluster-based sample retrieval for medical ner in clinical text. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 37–44, 2024.

[Ning *et al.*, 2023] Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. Skeleton-of-thought: Large language models can do parallel decoding. *Proceedings ENLSP-III*, 2023.

[Pu *et al.*, 2024] Hongxu Pu, Xincong Yang, Jing Li, and Runhao Guo. Autorepo: A general framework for multimodal llm-based automated construction reporting. *Expert Systems with Applications*, 255:124601, 2024.

[Steinstraeter *et al.*, 2020] M. Steinstraeter, J. Buberger, and D. Trifonov. Battery and heating data in real driving cycles, 2020.

[Tayarani *et al.*, 2023] Hanif Tayarani, Trisha V. Ramadoss, Vaishnavi Karanam, Gil Tal, and Christopher Nitta. Forecasting battery electric vehicle charging behavior: A deep learning approach equipped with micro-clustering and smote techniques. *arXiv preprint arXiv:2307.10588*, 2023.

[Wang *et al.*, 2024] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[Wang *et al.*, 2025] Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.