Towards the Adoption of Large Language Models for Interactive Digital Twin in Battery Management Systems

Judith Nkechinyere Njoku *, Cosmas Ifeanyi Nwakanma†, Jae-Min Lee *, and Dong-Seong Kim *
* Department of IT Convergence Engineering, Kumoh National Institute of Technology, 39177 South Korea
† Department of Computer Science and Electrical Engineering, West Virginia University, USA.

2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN) | 979-8-3315-2487-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICUFN65838.2025.11169958

***Abstract*—In this paper, we propose a novel digital twin assistant for battery management systems (BMS) to interpret the insights the digital twin provides. Our proposed digital twin assistant is based on a FLAN-T5 large language model (LLM), fine-tuned using a low-rank adaptation parameter-efficient fine-tuning approach for more domain-specific responses. The model was trained on a dataset specifically derived from a NASA battery dataset and was tailored to provide prompts and responses for training and testing the FLAN-T5 model. Experiment results show, the fine-tuned FLAN-T5 model achieves a rogue1, rogue2, rogueL, and rogueLsum scores of 0.400, 0.152, 0.286, and 0.292 respectively. These results highlight the potential of the proposed approach in ensuring a more explainable and user-friendly digital twin for BMS.**

***Index Terms*—digital twin, battery management systems, large language models**

I. Introduction

The efficient management of batteries is critical in enhancing battery systems' safety, performance, and lifespan [1], [2]. With the advent of the Internet of Things (IoT) and the integration of digital twins, battery management systems (BMS) have undergone a significant transformation. Digital twins, virtual replicas of physical systems, utilize IoT-generated sensor data, artificial intelligence algorithms, and sometimes physics models to provide real-time monitoring, diagnostics, and predictive maintenance capabilities [3]. This integration has improved the BMS field and addressed the complexities related to batteries' intricate physical and chemical processes [4].

Numerous studies have successfully developed fully functional digital twins for BMS. For instance, cloud-based digital twin systems were designed for BMS in [1] and [5]. Similarly, in [6], a digital twin model was also developed with a focus on state estimation algorithms. These two studies presented extensive results that highlighted the successes of their solutions. Despite these successes, some challenges with digital twins may need to be solved. Firstly, digital twins generate vast quantities of data that capture the complex behavior of batteries [7]. This data can often be high-dimensional and require substantial domain knowledge for interpretation. Secondly, the technical nature of digital twin outputs makes them mainly inaccessible to non-expert users who need to make informed decisions based on this data [8]. Thirdly, BMS's dynamic and real-time nature requires quick data interpretation for timely decision-making. The complexity of digital twin outputs can hinder this. Lastly, translating the predictive outputs from digital twins into understandable and actionable insights adds another layer of complexity.

The recent rapid growth in large language models (LLMs) has profoundly impacted the research landscape across various domains [9]. This impact can also be introduced to the digital twin domain. A few studies have already discussed the potential of utilizing LLMs in digital twins [10]. LLMs are exceptionally large deep learning (DL) models trained on vast amounts of data, enabling them to understand and generate human language [11], [12]. These models have demonstrated impressive performance in generating human-like text, making them valuable assets for tasks that require complex reasoning and natural language understanding [13]. LLMs offer a new avenue for enhancing the interactivity and user-friendliness of BMS through *digital twin assistants*. These assistants can revolutionize the decision-making processes and predictive analyses by converting complex digital twin insights into straightforward, actionable information. This evolution is essential, as existing literature primarily focuses on the technological and functional aspects of digital twins and IoTs in BMS, with little attention given to the user interaction layer, where LLMs can play a transformative role.

In this paper, we propose an innovative digital twin assistant to interpret the outputs of digital twins in BMS and serve as a liaison between the data generated by the digital twin and the actionable insights required by the end users. To achieve this, we present a two-stage process that utilizes the outputs of a machine learning model as input to a FLAN-T5 large language model. Initially, we developed a custom dataset composed of prompts and responses based on the battery dataset. Next, we use the generated dataset to train a FLAN-T5 model. Finally, we evaluate the entire system by having an ML model predict the battery capacity and then feeding the predicted capacity into the LLM along with prompts to provide an adequate response. Our main contributions are summarized as follows:

1) We generate prompts and responses about the state of a battery based on the battery dataset.

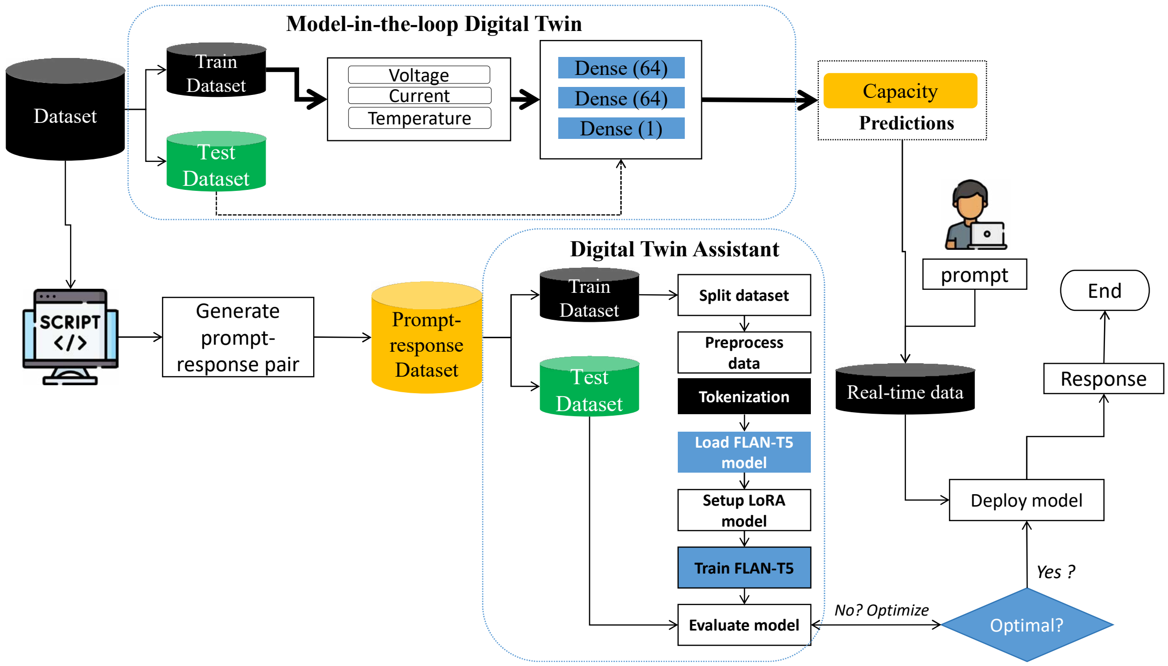

Fig. 1. Overview of our proposed approach. A model-in-the-loop digital twin predicts the capacity of the battery using the NASA dataset. This dataset is also used to generate a dataset comprised of pairs of prompts and responses. This new dataset is then used to train a FLAN-T5 model, fine-tuned using the LoRA approach. The electric vehicle user can interact with the digital twin assistant to obtain explainable analytics about the vehicle capacity.

2) We train a large FLAN-T5 language model to predict battery analytics based on the expected battery capacity and some prompts.
3) We apply parameter-efficient fine-tuning to the FLAN-T5 large language model.
4) We compare the performance of the optimized FLAN-T5 model and the expected response. Results highlight the applicability of the proposed method.

To our knowledge, this is the first study on applying LLMs for digital twin assistants for BMS.

II. Proposed System

This section introduces the overall system flow, including (i) problem formulation, (ii) generation of the dataset, and (iii) fine-tuning of the FLAN-T5 LLM. Fig. 1 illustrates the system model for the proposed digital twin assistant.

A. Problem formulation

We approach the development of a *digital twin assistant* for BMS with a two-fold strategy: First, we aim to predict the battery capacity C_{pred}, based on input features $\mathbf{X}$, such as voltage, current, and temperature, represented by the function:

$$C_{\text{pred}} = f_{\text{DNN}}(\mathbf{X}). \tag{1}$$

where f_{DNN} denotes the DNN model.

Second, we generate insightful advisories using the FLAN-T5 LLM based on the predicted capacity and prompts. Upon obtaining the predicted capacity, the FLAN-T5 model formulates an advisory A considering the capacity and provided prompts I.

$$A = g_{\text{LLM}}(C_{\text{pred}}, I). \tag{2}$$

The objective is to minimize the negative log-likelihood loss. The loss for a single output token, where the model predicts a probability distribution over the entire vocabulary, is defined as:

$$L = -\sum_{c=1}^{V} y_c \log(p_c), \tag{3}$$

Where V represents the vocabulary size, y_c denotes the one-hot encoded vector of the actual token, with 1 for the true token and 0 for all other tokens. p_c is the predicted probability for each token in the vocabulary. The sum is taken over all tokens in the vocabulary; however, due to the one-hot encoding, only the term corresponding to the true token contributes to the sum. For an entire sequence of tokens, the total loss is the sum of the individual token losses.

$$L_{\text{total}} = \sum_{i=1}^{N} L_i \tag{4}$$

where N is the number of tokens in the target sequence and L_i represents the loss associated with the i-th token in the sequence.

During the training process, the FLAN-T5 model aims to minimize this total loss, thereby adjusting its parameters to enhance the prediction accuracy of the output sequences. This optimization process enables the model to generate text sequences that are not only coherent but also contextually relevant to the given input prompts.

B. Generation of Dataset

A dataset of prompt-response pairs was generated, serving as the basis for training the LLM. These pairs capture the intricate interactions between various battery parameters and their implications, enabling the LLM to generate contextually appropriate advisories.

Algorithm 1 provides a guide for generating these pairs from a given dataset. Fig. 2 also illustrates the data generation process and an example of an interaction between an electric vehicle user and the digital twin assistant based on the prompts-response pair.

Algorithm 1: Generate Prompt-Response Pairs

Inputs: DataFrame D, Start ID $startID$
Output: Prompt-response pairs $Pairs$
function GENERATEPAIRS(D, $startID$) **forall** *rows* r *in* D **do**
    end
    $i \in \{1, 2\}$ ▷ Generate two pairs per row
    Generate $prompt_i$ based on r with focus variant for $i = 2$
    Generate $response_i$ based on $r.Capacity$
    $Pairs \leftarrow Pairs \cup \{(startID, prompt_i, response_i)\}$
    $startID \leftarrow startID + 1$
    return $Pairs$
end function
$Pairs \leftarrow$ GENERATEPAIRS(D, 1)
Save $Pairs$ to a CSV file

1) Generating prompts: The generation of prompts ($prompt1$ and $prompt2$) was designed to reflect the diverse scenarios encountered within BMS, based on the NASA battery dataset for battery $B0005$ [14]. By incorporating parameters such as $r.type$, $r.Battery$, and $r.id\_cycle$, the prompts simulate real-world conditions under which battery advisories are sought. This enhances the language model's training and ensures its generated advisories are rooted in practical, operational realities.

2) Response formulation: The corresponding responses ($response1$ and $response2$) are directly tied to the battery's capacity ($r.Capacity$), a crucial metric indicative of the battery's health and performance. Through these responses, the LLM learns to associate specific battery states and operational

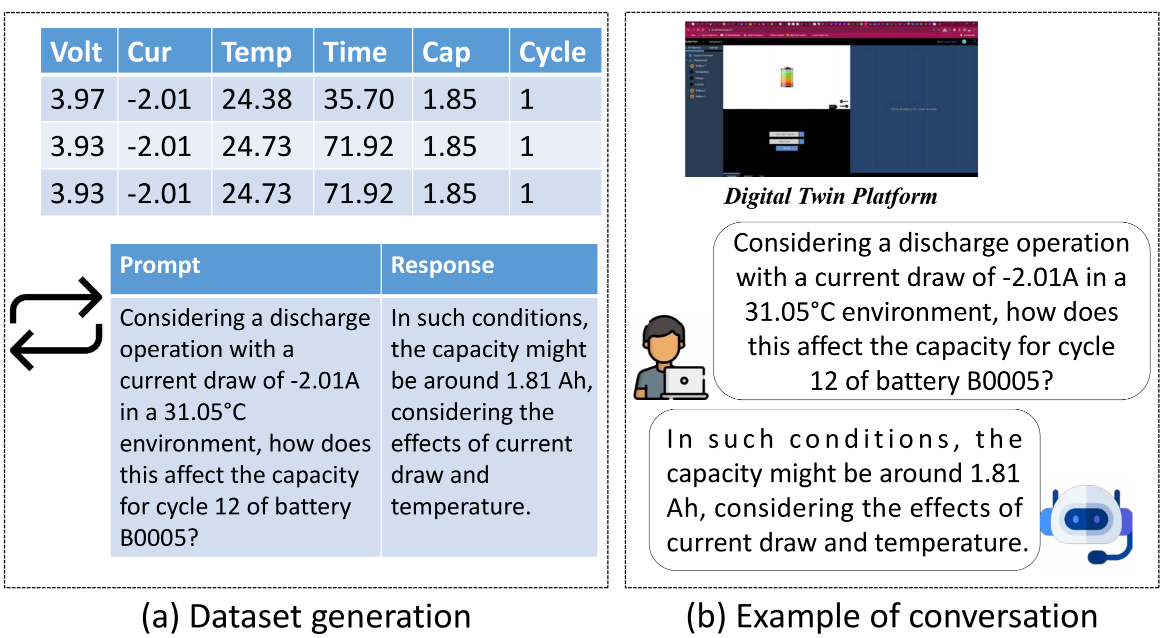

Fig. 2. (a) Generation of prompt and response pair using battery dataset. (b) Example of conversation between electric vehicle user and digital twin assistant.

conditions with their potential impact on capacity.

C. Development and fine-tuning of FLAN-T5

The FLAN-T5 is an open-source, sequence-to-sequence LLM published by Google in 2022 and trained on 1000 different tasks. The FLAN represents *Fine-tuned language Net*, and T5 represents *Text-to-Text Transfer Transformer*. The FLAN-T5 model is based on an encoder-decoder architecture as illustrated in Fig. 3.

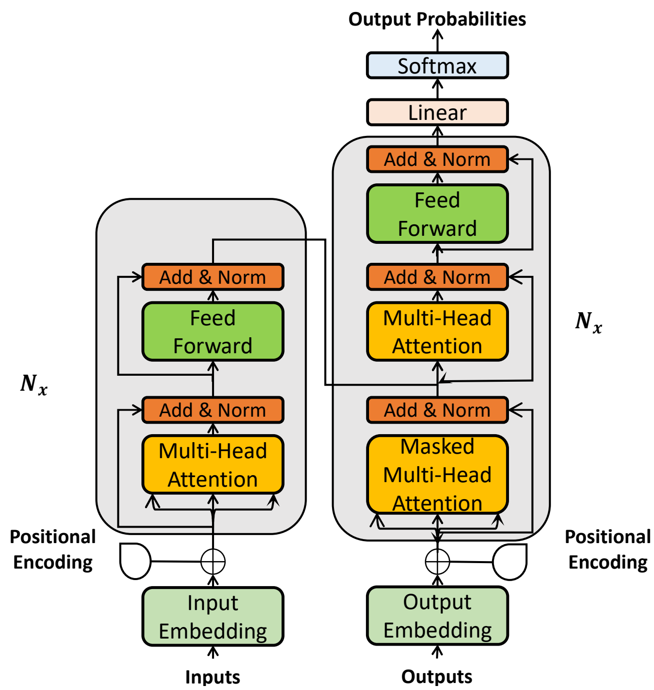

Fig. 3. Illustration of the FLAN-T5 model architecture

1) Fine-tuning the LLM: To transform the FLAN-T5 language model from a jack-of-all-trades to a domain-specific expert, we fine-tuned it using Parameter-efficient fine-tuning (PEFT). PEFT is an alternative approach that only updates a subset of the model weights. This is achieved by freezing all language model weights, adding a small number of new parameters, and updating only these new parameters. The Low-Rank Adaptation (LoRA) method of PEFT was adopted in this paper. LoRA is a method that minimizes the number of parameters that need to be fine-tuned by freezing the original model parameters and introducing a pair of low-rank matrix decomposition matrices alongside the frozen weights. The dimensions of the smaller matrices are set so that their product has the exact dimensions as the weight they are modifying. The original weights of the LLM remain static, while the smaller matrices are trained using supervised learning. During inference, the two low-rank matrices are multiplied to produce a matrix with the same dimensions as the frozen weights. This matrix is then added to the original weights and replaced with the updated values in the model. LoRA introduces a structured update to the weight matrices in the Transformer's attention and feed-forward layers, enabling parameter-efficient learning.

Given a weight matrix W in a transformer layer, LoRA decomposes it as $W = W_{\text{base}} + \Delta W$, where W_{base} represents the pre-trained weights, and $\Delta W = BA$ captures the low-rank updates introduced during fine-tuning, with B and A being trainable matrices. The loss function optimized during fine-tuning remains the same. The critical distinction in fine-tuning with LoRa lies in the gradient updates, primarily applied to the low-rank matrices B and A rather than directly to the full-weight matrix W. This approach allows for maintaining the foundational knowledge captured by W_{base} while efficiently adapting the model to the nuances of the target task. The working principles of LoRA is illustrated in Fig. 4.

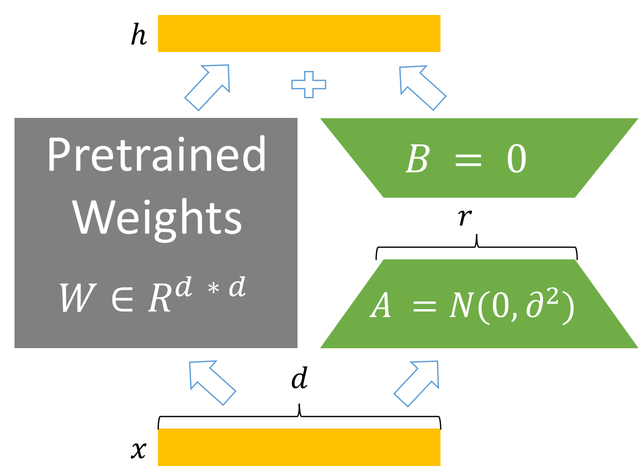

Fig. 4. How LoRA works for fine-tuning LLMs

2) Prompting the LLM: A standard language model method is by prompting, such as zero-shot and few-shot prompting. Zero-shot prompting refers to a scenario in which a language model is tested on a task it has not encountered before and relies on pre-trained knowledge to make predictions. In the case of few-shot prompting, the language model is given a small number of examples, allowing it to explore them better to understand the task structure and prompt. In this paper, both prompting approaches were employed.

3) Implementation approach: The text inputs were converted to token IDs using a Transformer Tokenizer to train the model. The dataset was split in a 90:10:10 ratio before preprocessing using abstractive summarization, a text2text generation task. Afterward, LoRA fine-tuning was applied to the model.

III. Experiments

A. Evaluation Dataset and Metrics Selection

The digital twin LLM model was trained and evaluated based on the generated dataset. $8,934$ data samples were generated

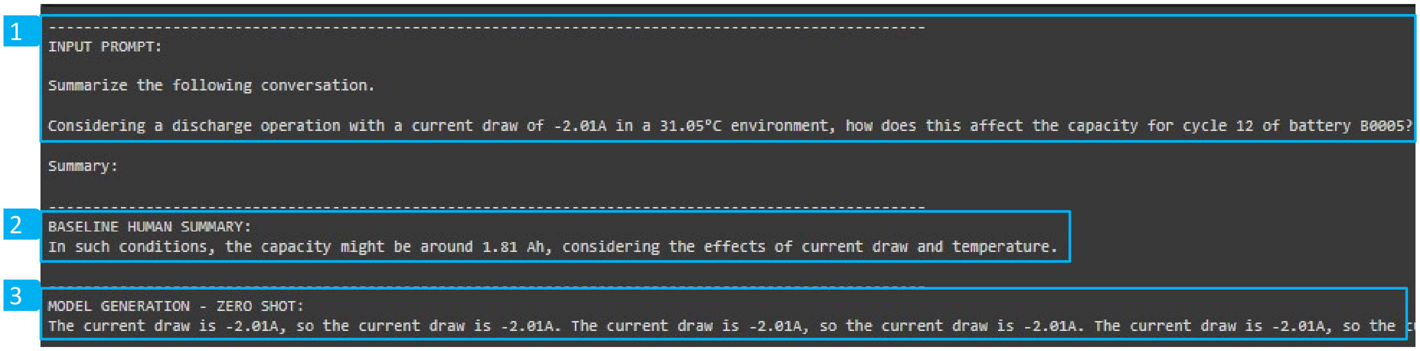

Fig. 5. Result of zero-shot prompting of the base FLAN-T5 model

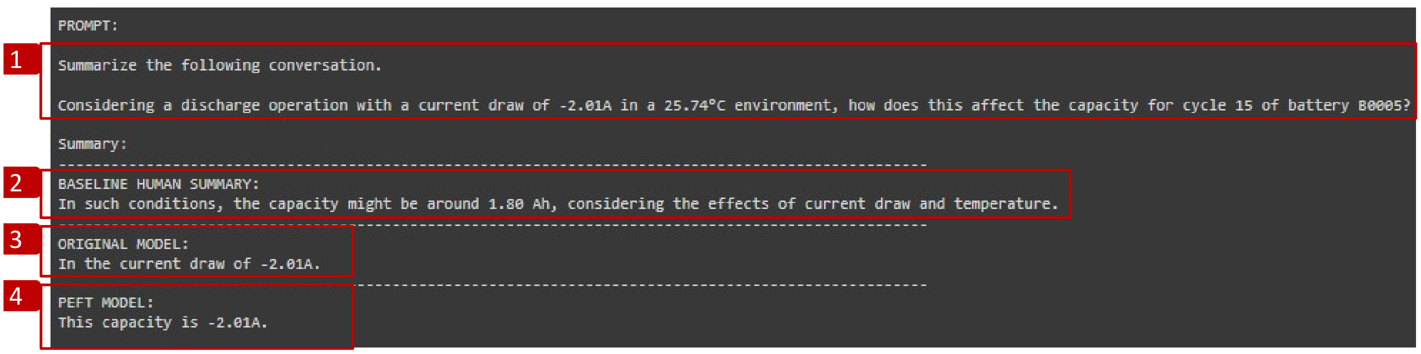

Fig. 6. Result of few-shot prompting of the base model and the LoRA fine-tuned model

from the battery dataset. The Rogue metrics measure the model's effectiveness and how well the digital twin assistant understands the prompts. Rogue-1 captures unigrams and measures how well the generated responses match the reference response. Rogue-2 captures bigrams by breaking each sentence into pairs of words, taking the number of overlapping bigrams, and dividing by the total number of bigrams in reference response. Rogue-L captures the longest common sequence and identifies the longest sequence of words that appear in the same order in both reference and digital twin LLM response. Rogue-Lsum applies the Rogue-L metric at the sentence level and aggregates all result for the final score.

B. Result and Analysis

The system was simulated in PyTorch in Google Colaboratory. The dataset was first saved in Google Drive before importing into the simulation environment. The FLAN-T5 model was trained in 50 epochs for an automated batch size that computes the batch size possible, a learning rate of 10×10^{-4}, a weight decay of 0.01, and for 50 logging steps. For finetuning, a rank of 32 was employed for the low-rank matrices, and $lora\_alpha$ was set to 32 with the target modules set to the query and key layers. A $lora\_dropout$ of 0.05 was used.

In the first experiment, we conducted zero-shot prompting on the original FLAN-T5 model, and compared the response with the baseline human response. The result of this experiment is represented in Fig. 5. The prompt was designed to query the model about other battery parameters and environmental conditions with an expected response related to the battery's capacity. As highlighted by the third highlight, the model response is not coherent and does not provide insight.

In the second experiment, we conducted few-shot prompting on the fine-tuned FLAN-T5 model, and compared the result to the base model. The result of this experiment is represented in Fig. 6. As highlighted, the fine-tuned model provided a straightforward and coherent response with errors in the capacity prediction.

TABLE I
EXPERIMENTAL RESULTS

Model	rogue1	rogue2	rogueL	rogueLsum
FLAN-T5	0.400	0.152	0.286	0.285
LoRA-FLAN-T5	0.319	0.173	0.298	0.292

We also evaluated the model based on the previously defined rogue1, rogue2, rogueL, and rogueLsum metrics. We conducted this evaluation using the first ten responses in the test dataset. These results, as illustrated in Table I, highlight the potential of LLMs in interpreting digital twin predictions in BMS.

IV. Conclusions

This paper proposes a novel digital twin assistant for BMS that utilizes LLMs. We fine-tune a FLAN-T5 LLM using the LoRA approach to enhance the domain-specificity of its responses. Our findings emphasize the potential of LLMs in interpreting digital twin insights. Future studies will focus on developing fully functional digital twins for BMS enabled by LLMs. These studies will also explore the viability of other LLMs and optimize them for better performance.

Acknowledgments

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government(MSIT) (IITP-2025-RS-2020-II201612, 33%) and by Priority Research Centers Program through the NRF funded by the MEST(2018R1A6A1A03024003, 33%) and by the MSIT, Korea, under the ITRC support program(IITP-2025-RS-2024-00438430, 34%)

References

[1] Y. Wang, R. Xu, C. Zhou, X. Kang, and Z. Chen, "Digital twin and cloud-side-end collaboration for intelligent battery management system," *Journal of Manufacturing Systems*, vol. 62, pp. 124–134, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612521002284

[2] J. N. Njoku, E. C. Nkoro, R. M. Medina, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Leveraging digital twin technology for battery management: A case study review," *IEEE Access*, vol. 13, pp. 21 382–21 412, 2025.

[3] H. Luo, L. Cheng, L. Yang, X. Zhao, R. Liao, and Y. Zhang, "A novel approach to building digital twin transformers by combining virtual-real sensing: An example of degree of polymerization distribution," *Measurement*, vol. 222, p. 113714, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224123012782

[4] J. N. Njoku, C. I. Nwakanma, and D.-S. Kim, "Explainable data-driven digital twins for predicting battery states in electric vehicles," *IEEE Access*, vol. 12, pp. 83 480–83 501, 2024.

[5] W. Li, M. Rentemeister, J. Badeda, D. Jöst, D. Schulte, and D. U. Sauer, "Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation," *Journal of Energy Storage*, vol. 30, p. 101557, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352152X20308495

[6] M. Zhou, L. Bai, J. Lei, Y. Wang, and H. Li, "A digital twin model for battery management systems: Concepts, algorithms, and platforms," in *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, J. Yao, Y. Xiao, P. You, and G. Sun, Eds. Singapore: Springer Nature Singapore, 2022, pp. 1165–1176.

[7] J. N. Njoku, C. I. Nwakanma, G. C. Amaizu, and D.-S. Kim, "Prospects and challenges of metaverse application in data-driven intelligent transportation systems," *IET Intelligent Transport Systems*, vol. 17, no. 1, pp. 1–21, 2023. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/itr2.12252

[8] J. N. Njoku, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Model comparison and selection for battery digital twin development using pybamm," in *The 33rd Joint Conference on Communication and Information (JCCI)*, 2023.

[9] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 4, pp. 1413–1430, 2024.

[10] Y. Sun, Q. Zhang, J. Bao, Y. Lu, and S. Liu, "Empowering digital twins with large language models for global temporal feature learning," *Journal of Manufacturing Systems*, vol. 74, pp. 83–99, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612524000372

[11] Y. Sun, Y. Xu, Z. Xie, Y. Shu, and S. Du, "Gptsee: Enhancing moment retrieval and highlight detection via description-based similarity features," *IEEE Signal Processing Letters*, vol. 31, pp. 521–525, 2024.

[12] Z. Chi, H. Huang, L. Liu, Y. Bai, X. Gao, and X.-L. Mao, "Can pretrained english language models benefit non-english nlp systems in low-resource scenarios?" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1061–1074, 2024.

[13] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush, "Interactive and visual prompt engineering for ad-hoc task adaptation with large language models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1146–1156, 2023.

[14] B. Saha and K. Goebel, "Battery Data Set," NASA Ames Prognostics Data Repository, 2007, nASA Ames Research Center, Moffett Field, CA. [Online]. Available: http://ti.arc.nasa.gov/project/prognostic-data-repository