2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS) | 979-8-3315-6599-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/MASS66014.2025.00111

PA-JJAMA: An LLM Based Intrusion Detection System for CAN Bus Networks

Joshua Quintano
School of Engineering and Computer Science
Oakland University
Rochester, United States
jquintano@oakland.edu

Yao Qiang
School of Engineering and Computer Science
Oakland University
Rochester, United States
qiang@oakland.edu

Huirong Fu
School of Engineering and Computer Science
Oakland University
Rochester, United States
fu@oakland.edu

***Abstract*—Recent studies have demonstrated significant success in detecting attacks on the Controller Area Network (CAN) bus network using machine learning and deep learning models, including convolutional neural networks and transformer-based architectures. Building on this foundation, our work investigates the use of large language models (LLMs) not only for intrusion detection but also for providing interpretable explanations of their decisions. We fine-tuned three LLMs, i.e., SecureBERT, LLaMA-2, and LLaMA-3, for intrusion detection on CAN bus data. Among them, LLaMA-3 delivered the best results, achieving SOTA performance on the Car-Hacking dataset. Beyond attack classification, we evaluated LLaMA-3's ability to generate reasoning for its decisions through zero-shot prompting. The model successfully articulated its rationale, particularly for Denial-of-Service (DoS) attacks, demonstrating strong potential for explainability in intrusion detection systems. These findings highlight the potential of LLMs to serve as a highly accurate intrusion detection system while simultaneously providing interpretable explanations, thereby enhancing the investigative capabilities of cybersecurity professionals.**

***Index Terms*—CAN Bus Network, Intrusion Detection System, Large Language Models**

I. Introduction

The Controller Area Network (CAN) bus enables fast and efficient communication between Electronic Control Units (ECUs) within a vehicle's internal network and has remained an industry standard since its introduction [4]. Its widespread adoption is driven by simplicity, low cost, and the ability to support real-time communication among multiple ECUs, which are the key requirements for modern automotive systems. While newer protocols like CAN-FD and automotive Ethernet offer improved security and higher bandwidth, CAN remains the dominant choice for in-vehicle communication due to its lightweight design and cost efficiency. However, the protocol was designed with performance rather than security in mind and lacks essential features such as authentication and encryption. This vulnerability leaves CAN networks highly susceptible to a wide range of cyberattacks. Consequently, extensive research has focused on enhancing CAN security, with Intrusion Detection Systems (IDS) emerging as one of the most practical solutions. IDS approaches provide robust protection while maintaining CAN's real-time performance by passively monitoring network traffic without disrupting its broadcast-based architecture.

With the rapid advancement of artificial intelligence (AI), recent studies have increasingly explored the integration of deep learning techniques into IDS to enhance the security of CAN bus networks. Unlike traditional IDS approaches, which often rely on pre-defined rules or signatures, deep learning models offer the ability to identify complex patterns and potentially detect zero-day attacks or previously unseen malicious activity. Among these techniques, large language models (LLMs) have emerged as a promising solution due to their capacity to process vast amounts of data efficiently and leverage extensive pre-training on diverse datasets. This enables LLMs to capture nuanced relationships and anomalies, making them well-suited for identifying sophisticated attack vectors within automotive communication systems.

LLMs have been extensively studied for their ability to detect attacks within CAN bus networks, achieving notable success in distinguishing benign messages from malicious ones and accurately classifying attacks into specific types [1]. Building on this foundation, our work investigates the use of LLMs not only for message and attack-type classification but also for their capability to perform zero-shot reasoning. Specifically, we explore how LLMs can articulate the underlying patterns and rationale behind their classifications in response to user prompts, thereby enhancing explainability in intrusion detection. In our work, we used large language models to not only classify messages as being malicious or benign, but also began exploring the use of these large language models as tools that can assist professionals by providing explanations behind their classifications.

2155-6814/25/$31.00 ©2025 IEEE
DOI 10.1109/MASS66014.2025.00111

II. Related Works

We review prior research on deep learning-based CAN bus intrusion detection, summarize key approaches, and discuss the distinctions of our PA-JJAMA method.

Fu et al. introduced a BERT based model that leveraged a semantic extractor to train the model to be able to recognize the semantics of multiple different protocols, allowing it to detect attacks in multiple types of networks, such as IoT and the CAN bu [2]. The Car Hacking dataset was modified to include only the ID and payload data of the CAN messages. They sampled 10,000 messages from the normal messages and each type of attack message from the dataset. The model performed extremely well, with over 99.9 percent in every evaluation metric.

In a more CAN bus focused work, Nwafor et al. propose the CANBERT model[3]. They pre-trained the BERT model on the OTIDS [8] dataset to teach the model how to recognize CAN bus semantics. Once pre-trained, the model was then fine-tuned for CAN bus message classification. The dataset was divided three ways, 64% for training, 20% for validation, and 16% for testing. The model was said to have performed with near perfect detection but further specifics were not given.

In the work produced by Rai et al. they compare the performance of four deep learning models in their ability to detect attack in an in-vehicle network. The four models used were Long Short-Term Memory (LTSM), Gated Recurrent Unit (GRU), Bi-directional LSTM, and the pre-trained VGG-16 model. The work used the Car Hacking Dataset, OTIDS, and Survival Analysis datasets individually and also created a merged dataset of the three that consisted of over 6,000,000 total messages with over 1,000,000 attack messages. The dataset was split 80% for training with 20% for testing. The models were evaluated using accuracy, precision, recall, and F1-score. On the Car Hacking dataset VGG-16 performed the best, maintaining over 90 percent points in every evaluation metric. It performed worse in every evaluation metric than the language model based models discussed however.

Touvron et al. released Llama-2, an open-source models featuring different versions with parameter sizes of 7 billion, 12 billion, and 70 biliio [10]. The model waS trained on 2 trillion tokens on factual sources to attempt to increase the model's knowledge and reduce its hallucinations. We use this model as one of the benchmarks for LLM perfomance in intrusion detection. LLaMA-2 would be succeeded with the introduction of Llama-3 in 2024 by Dubey et al. LLaMA-3 would be even larger than Llama 2, featuring models with parameter sizes of 8 billion, 70 billion, and 405 billion [11]. The model was trained on 15.6 trillion tokens. This model is used as another benchmark for the evaluation of LLMs in CAN bus IDS architecture.

Agahei et al. introduced secureBERT in [7]. The model is based on the BERT architecture, and has been pretrained on cybersecurity related text and articles to teach it domain specific vernacular, with this corpora totaling over 2.2 million documents [12]. This model provides an encoder-based foil to benchmark language model performance in intrusion detection systems as a opposed to the two decoder-leveraging models of LLaMA-2 and LLaMA-3.

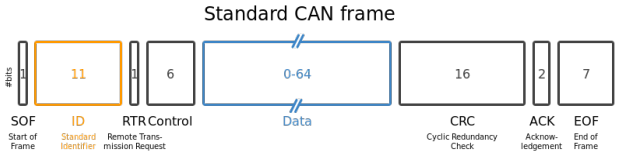

Fig. 1: CAN Message Frame [15]

Although these works demonstrate the effectiveness of deep learning and language models for the detection of CAN bus attacks, their focus has mainly been on classification. None have explored the use of LLMs for explanatory reasoning, an essential capability to improve transparency and aid cybersecurity professionals in forensic analysis. Our work addresses this gap by incorporating zero-shot reasoning to explain model decisions in addition to achieving high classification performance.

III. Background

A. CAN Bus Architecture

The CAN bus is a message-based serial communication protocol designed to enable high-speed data exchange within a vehicle's internal network. In a CAN bus architecture, ECUs, also referred to as nodes, communicate by broadcasting messages to the shared bus whenever it is available [5]. As a result, every ECU on the network can receive all transmitted messages. The bus is typically divided into two segments: a high-priority network and a low-priority network. The high-priority network handles critical communications such as steering assistance and acceleration control, whereas the low-priority network manages non-essential functions like window or sunroof operations. Message arbitration within the CAN bus is governed by a priority system, where the ECU with the lowest identifier is granted transmission priority.

Because the CAN bus protocol lacks authentication and encryption mechanisms, it is highly vulnerable to attackers with knowledge of its operation. Exploits can vary significantly in severity, from triggering minor actions such as opening or closing windows to executing critical commands like manipulating steering control while the vehicle is in motion [6].

B. SecureBERT

Transformer-based models form the foundation of most modern models for natural language processing (NLP). These architectures leverage encoders, decoders, and, most importantly, the self-attention mechanism combined with feed-forward neural networks to capture contextual relationships and generate coherent, human-readable text.

SecureBERT is a domain-specific language model tailored for cybersecurity applications and built on the RoBERTa

TABLE I: Car Hacking Dataset Statistics

Attack Type	Total Messages	Normal Messages	Injected Messages
DoS Attack	3,665,771	3,078,250	587,521
Fuzzy Attack	3,838,860	3,347,013	491,847
Gear Spoof	4,443,142	3,845,890	597,252
RPM Spoof	4,621,702	3,966,805	654,897

TABLE II: Balanced Car-Hacking Dataset Statistics

Attack Type	Total Messages	Normal Messages	Injected Messages
DoS Attack	1,175,042	587,521	587,521
Fuzzy Attack	983,694	491,847	491,847
Gear Spoof	1,194,504	597,252	597,252
RPM Spoof	1,309,794	654,897	654,897
Total	4,663,304	2,331,517	2,331,517

architecture [12]. It is pretrained on over 98,000 cybersecurity-related corpora and optimized for downstream tasks within this domain. The RoBERTa framework employs bidirectional encoding, enabling the model to process text in both directions across layers. [13] This bidirectional training strategy allows the model to develop a deeper contextual understanding of input sequences and effectively learn intricate patterns in the data. In our approach, CAN message data was tokenized into 11 tokens, with the final token representing the label that indicates whether the message is benign or malicious.

C. LLaMA Models

The LLaMA family of models consists of decoder-only models trained on vast amounts of publicly available open-source data. One of the key advantages of LLMs is their extensive pre-training, which enables strong generalization without requiring full retraining for most tasks. In this study, we fine-tuned LLaMA-2 and LLaMA-3 for two primary objectives: intrusion detection and explanatory response generation. To adapt these models for CAN bus intrusion detection, we transformed each CAN message into a prompt-answer format, where the prompt followed the structure ”Here is a CAN Bus Message:... Is this message malicious? If no answer ”No - Benign” otherwise ”Yes - Attack Type”” and the corresponding answer was either ”No - Benign” or ”Yes - Attack Type” where the attack type was replaced with the predicted attack type classified. The fine-tuned models were then evaluated based on their ability to produce correct responses, which were used to compute standard evaluation metrics. Additionally, we employed zero-shot prompting techniques to encourage the models to provide reasoning for their classifications, enabling insights into the patterns they recognized during decision-making.

IV. Data Pre-processing

A. Dataset

The Car Hacking Dataset developed by Song et al. [9] was utilized to fine-tune the three selected models for CAN bus intrution detection. This dataset contains over 12 million messages, including four injected attack types: Denial of Service, Fuzzy Attack, Gear Spoofing, and RPM Spoofing. To ensure format consistency, records with payload lengths shorter than eight bytes were padded with “00” in the data fields. Each attack-specific subset was individually evaluated using the three models. Subsequently, we constructed a balanced dataset by sampling all attack messages from each subset and an equal number of normal messages. An 80/20 split was applied for training and testing across all datasets. Additional details about the attack types are provided in Table I.

1) DoS Attack: Messages were injected with ”0000” as the ID every 0.3 milliseconds.

2) Fuzzy Attacks: Messages with completely random ID and Data values were injected every 0.5 milliseconds.

3) Gear & RPM Spoofing: Messages with ID values related to RPM and gear activity were injected every 1 millisecond. [9]

B. SecureBERT Data Pre-processing

For SecureBERT preprocessing, each CAN message was divided into the following fields: Timestamp, ID, Length, Data[0]–Data[7], and Status. The ID and Data fields were converted from hexadecimal to numeric values to enable tensor representation. In single-attack datasets, the Status field was mapped such that “R” (regular) was converted to 0 and “T” (attack) to 1, representing normal and malicious messages, respectively. For the combined dataset, each attack type was assigned a unique integer label as detailed in Table III.

C. LLaMA Model Data Pre-processing

For the LLaMA models, the Car Hacking Dataset was formatted such that each CAN bus message was embedded within a prompt asking whether the message was malicious. The corresponding answer began with “yes” or “no,” followed by a brief clarification indicating either the specific attack type or a benign status, as appropriate. For the Multi-Attack dataset, the four LLaMA preprocessed subsets were merged into a single dataset. Additionally, an alternative version was created by modifying the prompts to explicitly encourage zero-shot reasoning, enabling the model to provide explanations for its classifications.

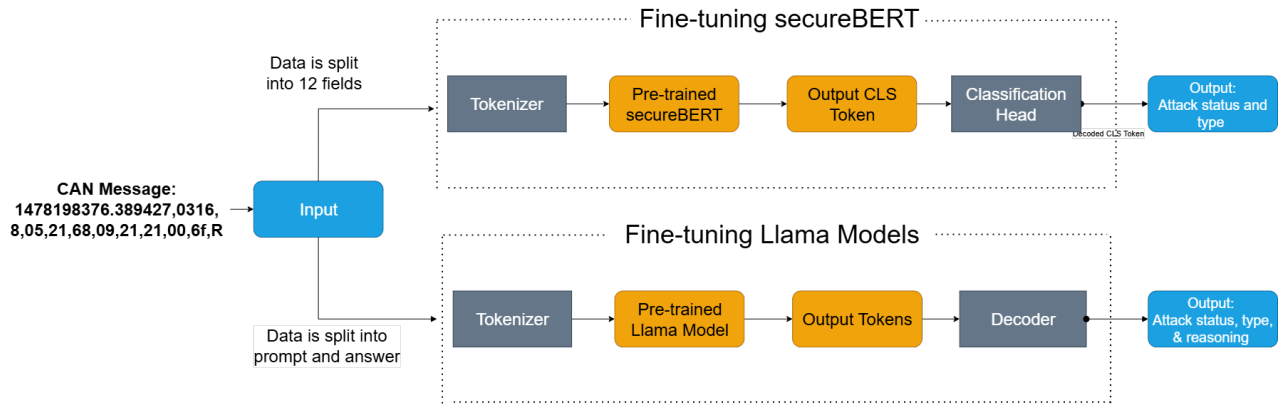

Fig. 2: SecureBERT Fine-tuning & LLaMA Model Fine-tuning

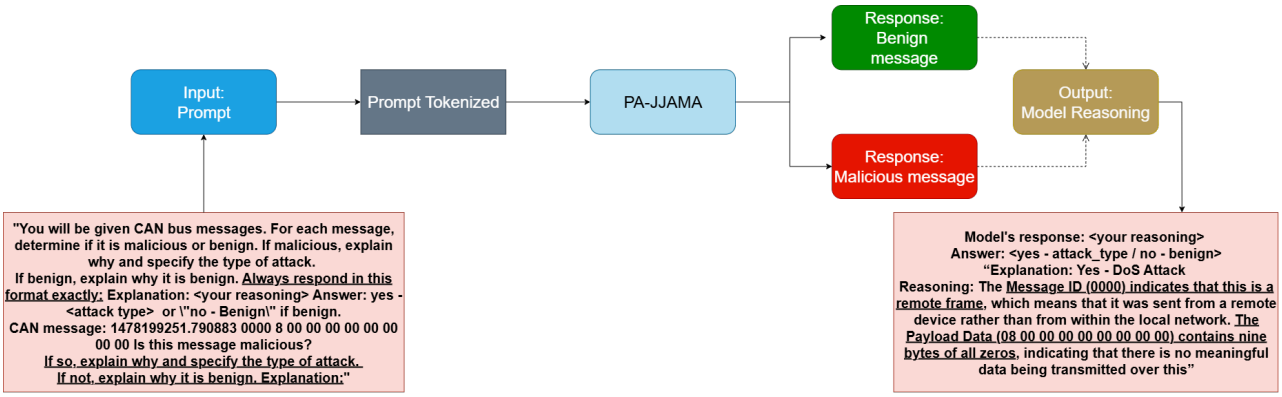

Fig. 3: PA-JJAMA Workflow

TABLE III: SecureBERT Message Labels

Attack Type	Label
Benign	0
DoS Attack	1
Fuzzy Attack	2
Gear Spoof	3
RPM Spoof	4

V. Method

A. SecureBERT Fine-tuning

For the single-attack-type datasets, the SecureBERT fine-tuning process involved training the top layer for a binary classification task, using malicious and benign as the labels. In contrast, for the combined dataset, the model was fine-tuned for multi-class classification, enabling it to both detect an attack and identify the specific attack type.

B. Llama Model Fine-tuning

The models are initially fine-tuned in text-classification mode to classify each message as malicious or benign and, when applicable, identify the specific attack type. This process was applied to both single-attack and combined-attack datasets. Subsequently, LLaMA-3 was further trained using prompts designed to encourage zero-shot reasoning based on patterns learned during fine-tuning. To accommodate hardware constraints and improve efficiency, we applied 4-bit quantization to reduce model precision and employed LoRA [14] to minimize the number of trainable parameters. These optimizations significantly reduced training time while maintaining model performance.

The PA-JJAMA architecture is based on Llama 3 and is trained in causal mode. CAN messages are input as a part of a prompt the promotes zero-shot reasoning. The model is only given the classification of the message and its attack type if malicious, no reasoning for the classification is given at any point during training.

VI. Results

A. Model Result Comparison

All three models demonstrated outstanding performance in distinguishing between benign and malicious CAN messages across both the individual attack-type datasets and the combined dataset.

Although SecureBERT achieved the lowest performance among the three models, its results remain impressive given its

TABLE IV: Model Performance on Individual Attack Datasets

Attack Type	SecureBERT			LLaMA 2			LLaMA 3		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
DoS	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Fuzzy	0.999996	1.000000	0.999970	0.999999	1.000000	0.999999	1.000000	1.000000	1.000000
Gear Spoof	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
RPM Spoof	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

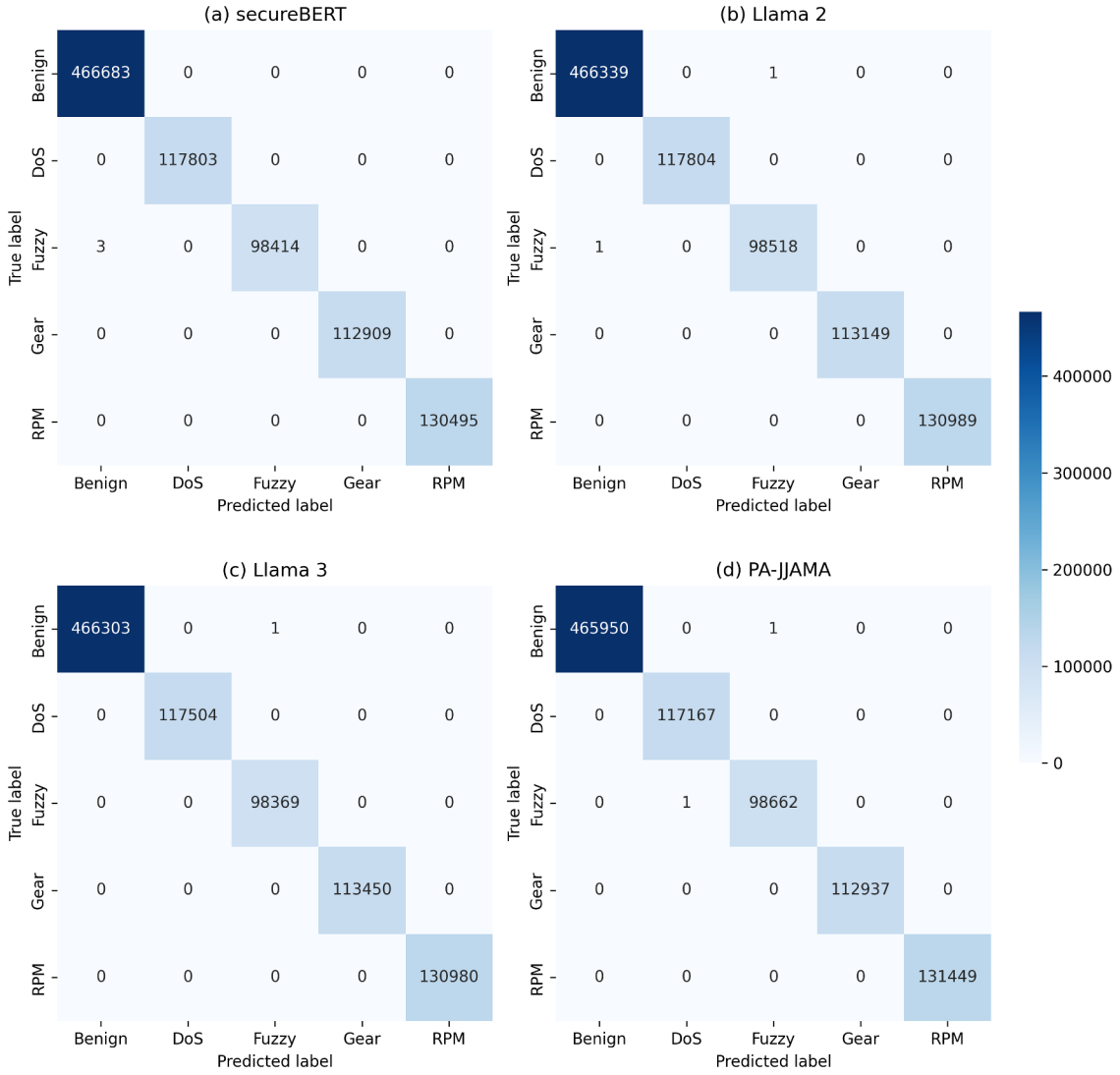

Fig. 4: Multi-Attack Dataset Performance

significantly smaller size, 110 million parameters compared to LLaMA-2's 7 billion and LLaMA-3's 8 billion. Across both single-attack and multi-attack datasets, SecureBERT misclassified only three messages. Its bidirectional encoding architecture clearly contributed to its ability to capture context and learn message patterns, resulting in strong overall performance in attack detection.

LLaMA-2 achieved strong results, misclassifying only two messages across all individual attack-type datasets. LLaMA-3 delivered even better performance, correctly labeling every message in the single-attack datasets and misclassifying just two messages in the combined dataset. Given

the difference in parameter sizes, this slight improvement by LLaMA-3 over LLaMA-2 aligns with expectations. While LLaMA-2 performed marginally worse on the multi-attack dataset, it still achieved near-perfect accuracy with only two errors. This performance gap may be attributed to LLaMA-3's larger parameter count, which likely enhanced its ability to capture subtle patterns within the data. Our proposed model, PA-JJAMA, performed comparably to LLaMA-2 but misclassified only two messages. Overall, all three models demonstrated exceptional accuracy, and the self-attention mechanisms of the LLaMA models appear to have played a critical role in learning complex token relationships and identifying patterns associated with malicious payloads.

B. Zero-Shot Reasoning with PA-JJAMA

Building on its strong performance in attack detection and classification, LLaMA-3 was further fine-tuned to develop PA-JJAMA. This model was designed not only to identify malicious messages and classify their attack types but also to provide reasoning for its decisions. To achieve this, we employed a zero-shot methodology, training the model with prompts that emphasized explanatory responses. However, during training, the model's outputs consisted primarily of a binary answer ("yes" or "no") followed by the classification label: Benign, DoS Attack, Fuzzy Attack, Gear Attack, or RPM Attack. To address this, we introduced an inference-time prompt that explicitly encouraged reasoning in addition to classification. The resulting causal version of the model maintained high accuracy, misclassifying only two messages in the multi-attack dataset.

Although the response incorrectly identified the DLC field as part of the data payload, its ability to provide a coherent explanation and accurately highlight the key features that informed its decision is highly promising, as shown in Figure 3.

VII. Conclusion

In this work, we evaluate the application of LLMs for intrusion detection and introduce PA-JJAMA, a model capable of not only identifying attacks but also providing reasoning for its classifications when prompted. While PA-JJAMA does not consistently use correct field names in its explanations, it successfully identifies key features within the data that indicate malicious activity. Further refinement of training prompts is needed to improve the accuracy and specificity of these explanations. Despite these limitations, PA-JJAMA demonstrates the potential of LLMs to extend beyond task automation and serve as valuable tools for cybersecurity professionals by offering both detection and interpretability.

References

[1] H. Kheddar, "Transformers and Large Language Models for Efficient Intrusion Detection Systems: A Comprehensive Survey," *arXiv*, Aug. 2024. [Online]. Available: https://doi.org/10.48550/arxiv.2408.07583

[2] M. Fu, P. Wang, M. Liu, Z. Zhang, and X. Zhou, "IoV-BERT-IDS: Hybrid network intrusion detection system in IoV using large language models," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 1–13, Jan. 2024, doi: 10.1109/tvt.2024.3402366.

[3] E. Nwafor and H. Olufowobi, "CANBERT: A language-based intrusion detection model for in-vehicle networks," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Nassau, Bahamas, Dec. 2022, pp. 294–299, doi: 10.1109/icmla55696.2022.00048.

[4] S. Sharma, "Understanding CAN Bus: A comprehensive guide," *Wevolver*, Nov. 7, 2023. [Online]. Available: https://www.wevolver.com/article/understanding-can-bus-a-comprehensive-guide

[5] M. Falch, "CAN Bus explained — a simple intro (2021)," *CSS Electronics*, Jan. 2025. [Online]. Available: https://www.csselectronics.com/pages/can-bus-simple-intro-tutorial

[6] A. Greenberg, "Hackers remotely kill a Jeep on the highway—with me in it," *WIRED*, Jul. 21, 2015. [Online]. Available: https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

[7] E. Aghaei, "SecureBERT: A domain-specific language model to represent cybersecurity textual data," *GitHub*, 2025. [Online]. Available: https://github.com/ehsanaghaei/SecureBERT (accessed Jul. 15, 2025).

[8] H. Lee, S. H. Jeong, and H. K. Kim, "OTIDS: A novel intrusion detection system for in-vehicle network by using remote frame," in *Proc. Annu. Conf. Privacy, Secur. Trust (PST)*, Calgary, AB, Canada, Aug. 2017, doi: 10.1109/pst.2017.00017.

[9] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Veh. Commun.*, vol. 19, p. 100198, Oct. 2019, doi: 10.1016/j.vehcom.2019.100198.

[10] T. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "LLaMA 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, Jul. 2023.

[11] 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*, Jul. 2024

[12] E. Aghaei, X. Niu, W. G. Shadid, and E. Al-Shaer, "SecureBERT: A Domain-Specific Language Model for Cybersecurity," in *Security and Privacy in Communication Networks: 18th EAI International Conference, SecureComm 2022, Virtual Event, October 2022, Proceedings* (Fengjun Li, Kaitai Liang, Zhiqiang Lin, Sokratis K. Katsikas, eds.), Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, vol. 462, pp. 39–56, Springer, Feb. 2023. doi:10.1007/978-3-031-25538-0_3

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[14] E. S. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models.," Jun. 2021.

[15] M. Falch, "CAN Bus Explained - A Simple Intro (2021)," CSS Electronics, Jan. 2025. https://www.csselectronics.com/pages/can-bus-simple-intro-tutorial