Towards a Hybrid LLM-Based Intrusion Detection System for Cyber-Physical Systems Applications

Mamdouh Muhammad[1][0009−0009−9685−8880], Abdelkader Magdy Shaaban[2][0000−0002−9159−8436], Reinhard German[1][0000−0000−9071−4802], and Loui Al Sardy[1][0000−0002−8461−5154]

[1] Computer Networks and Communication Systems, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
{mamdouh.muhammad, reinhard.german, loui.alsardy}@fau.de
https://www.cs7.tf.fau.de/

[2] Center for Digital Safety & Security, Austrian Institute of Technology, Vienna, Austria
abdelkader.shaaban@ait.ac.at
https://www.ait.ac.at/en/

Abstract. The increasing complexity of cyberattacks on the Cyber-Physical Systems (CPS) demands for advanced intrusion detection strategies capable of interpreting contextual threats. Conventional hybrid Intrusion Detection Systems (IDSs) suffers from outdated attack signature databases and limited attack insights. This paper proposes a conceptual framework for an advanced hybrid IDS that integrates Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG). Our framework combines signature-based and anomaly-based detection with an LLM-RAG threat analysis module to provide context-aware classification of network traffic events in the context of domain knowledge. We outline potential implementation challenges and propose preliminary mitigation strategies. Future work will focus on empirical validation through experimental evaluation. This framework demonstrates the viability of LLM-RAG-powered IDS for CPS cybersecurity.

Keywords: Society 5.0, Cyber-physical systems, Intrusion detection system, Large language model, Retrieval-augmented generation, Contextual reasoning, Cybersecurity

1 Introduction

Society 5.0 is a Japanese-origin concept to integrate technologies such as Big Data, Artificial Intelligence (AI), Internet of Things (IoT), and robotics into daily life, enabling a smart and human-centred society [1]. This concept extends into domains like Industry 5.0, Farming 5.0, smart health, smart mobility, and smart cities [2] and [3], emphasising collaboration between humans and machines

that expand human capabilities. Among these domains, Cyber-Physical Systems (CPS) form the technological backbone for smart infrastructures.

Applications such as Industrial Control System (ICS) and smart grids are two important domain-specific applications of CPS [4]. While ICS focuses on control and implementation of industrial processes, smart grids integrate Information Technology and Operational Technology (IT/OT), increasing sustainability and efficiency, while enabling two-way communication between smart grids components[5]. However, this integration introduces new attack surfaces, increasing the vulnerability of critical infrastructure to cyber threats.

In response, governments and organisations have introduced regulatory frameworks to guide and enforce cybersecurity standards. For example, the **BSI Act** (Germany) and its extension, the **IT Security Act 2.0** [6] mandate protection of critical infrastructure, including Intrusion Detection System (IDS) deployment [7], and [8]. Similarly, the **EU AI Act** [9] is considered the first regulation on AI that handles and addresses the risks of AI in four categories. For example, article 15 in the high-risk AI systems - like critical infrastructure [10] - chapter emphasises ensuring appropriate technical solutions to ensure cybersecurity of such systems [11]. Therefore, advanced hybrid IDSs are crucial in securing CPS, as conventional hybrid IDSs often struggle to detect zero-day attacks or adapt to evolving threats due to outdated datasets and limited contextual reasoning.

Large Language Models (LLMs) are pre-trained language models that have powerful and efficient capabilities in many Natural Language Processing (NLP) tasks [12]. Besides being used in text summarisation and generation, translation, and question answering, LLM's prominent capability is context reasoning, which leverages the contextual information to perform reasoning tasks [13]. Therefore, they offer a promising research area for enhancing IDS performance. However, applying LLMs as analysis modules raises several challenges, such as the risk of hallucinations [14], swinging performance [15], nondeterministic nature, and model interpretability and bias [16].

Due to the aforementioned limitations of the conventional hybrid IDS, such as context-blind nature and the limitation of LLM, such as hallucinations, in this paper, we propose a Hybrid LLM-based IDS (**HyLLM-IDS**) framework tailored for a CPS environment. Our framework combines conventional detection engines with an LLM-based RAG module to provide context-aware threat analysis and detection capabilities.

The remainder of the work is organised as follows: Sect. 2 discusses related work on IDS and LLMs in cybersecurity, Sect. 3 introduces the proposed methodology, Sect. 4 presents a discussion of the challenges and possible mitigations, and Sect. 5 outlines conclusions and future directions.

2 Related Work

This section reviews foundational and state-of-the-art techniques in intrusion detection. We first introduce the conventional signature-based and anomaly-based IDS approaches, highlighting their strengths and limitations in detecting both

known and novel threats. Subsequently, we examine the recent integration of LLMs into cybersecurity, focusing on their role in enhancing threat understanding, contextual reasoning, and explainability. Our aim is to complement these existing efforts with our proposed HyLLM-IDS framework and underline the unique contributions it brings to CPS security.

2.1 Signature-based IDS and Anomaly-based IDS

Signature-based IDS detect threats by matching predefined patterns or known attack signatures to detect anomalies. Although such IDSs have low false positive rates, they are ineffective against zero-day attacks, which do not match existing signatures. In contrast, anomaly-based IDS can detect unknown or zero-day attacks by identifying irregularities and variations from normal behaviour [17]. Although anomaly-based IDSs have low false negative rates, they rely on training the models on popular, mostly publicly available datasets like UNSWNB15 [18], CIDDS001 [19], CIC-IDS2018 , CIC-DDoS2019, CIC-IoT2023 [20], as Nguyen et al. investigated in [21]. However, reliance on these static and publicly available datasets limits the generalisability of trained models, where attacks are evolving and becoming more sophisticated.

Many researchers tried combining signature-based and anomaly-based IDSs to balance low false positive and low false negative rates, as in [22,23]. However, the problem in such approaches is that when training Machine Learning (ML) or Deep Learning (DL) models on such limited datasets still restricts the detection scope, making them ineffective against newly emerging threats.

2.2 LLMs in Cybersecurity

LLMs can analyse sequences of events and correlate them with known cybersecurity knowledge, potentially improving detection accuracy and providing interpretation for alerts.

In [24], Benabderrahmane et al. introduced Advanced Persistent Threats LLM (APT-LLM) as a novel embedding-based anomaly detection framework that incorporates autoencoders and LLMs (BERT, ALBERT, DistilBERT, and RoBERTA) to detect APTs. Although the authors claim that the results outperform other anomaly detection methods, their approach was used and tested on only DARPA provenance logs, which raises the question about the performance if other logs are used, like network or application logs. In addition, their frameworks are not designed for real-time use.

In [25], Ghosh et al. developed Common Vulnerabilities and Exposures LLM (CVE-LLM) as a system to assess vulnerabilities automatically. The authors trained the model on historical assessments of medical device vulnerabilities. In addition, they added data from CVEs and Common Weakness Enumerations (CWEs) to enrich and expand the model training data. The authors also mentioned that domain adaptation increased their model accuracy in comparison to other models (e.g., Mistral-7B, LLama2-7B) in two assessments, Common Vulnerability Scoring System (CVSS) Vectors and Vulnerability Exploitability

eXchange category (VEXCategory). Their domain-specific model used a large set of annotated historical assessments, as mentioned before. Such limitations pose challenges in generalising its performance to other domains, and if there is not enough available annotated data.

As logs are important in the cybersecurity realm, in [26], Zhong et al. introduced LogParser-LLM as an efficient log parsing LLM. Their model uses ChatGPT (version gpt-3.5-turbo-0301) and GPT-4 (version gpt-4-0613) for template extraction. Evaluated on two benchmarks LogHub and LogPub, and the models showed a high F1 score for grouping accuracy and for parsing accuracy, outperforming other log parsers. However, the us of OpenAI's commercial APIs poses potential cost and accessibility issues.

Lastly, in [27] Song et al. introduced Audit-LLM as a multi-agent LLM framework that analyses auditing logs to identify threats. The models consist of three agents: the decomposer that applies Chain-of-Thought (CoT) reasoning to deconstruct the problem, a tool builder that generates mini-tools to handle subtasks, and an executor to executes these tools to conclude. They tested the model on three Insider Threat Detection (ITD) datasets showing its validity in improving the generated explanations. Using multiple agents can serve many tasks within the LLM's overall mission, but this makes it computationally expensive, and using three ITD-oriented datasets can raise questions about the model's scalability and generalisation.

3 Proposed Methodology

The proposed HyLLM-IDS framework, illustrated in Fig. 1, consists of three interconnected components: (i) a parallel IDS detection module, (ii) a context-aware threat analysis engine powered by an LLM, and (iii) a RAG system for improving threat intelligence. Together, these modules enable robust, multi-stage detection that combines conventional and adaptive cybersecurity techniques. An overview of each component is provided in the following subsections:

3.1 IDS Detection Module

The IDS Detection Module (*step 2*) integrates a parallel deployment of signature-based IDS and anomaly-based IDS. Every network flow or packet is simultaneously analysed by both subsystems. If either detects an anomaly flag, the corresponding data is forwarded to the LLM for contextual analysis. This parallel setup maximises recall for zero-day attacks. The following describes each IDS:

Signature-based IDS. (*step 3*) For the signature-based component *Suricata* [28] was selected due to its high-quality deep packet inspection capabilities, support of multi-threading, flexible log export in JSON format, and functions as IDS/Intrusion Prevention System (IPS)/Network Security Monitoring (NSM). Additionally, it is open-source, has moderate installation complexity, and can be leveraged for advanced rule customisation using Lua scripting.

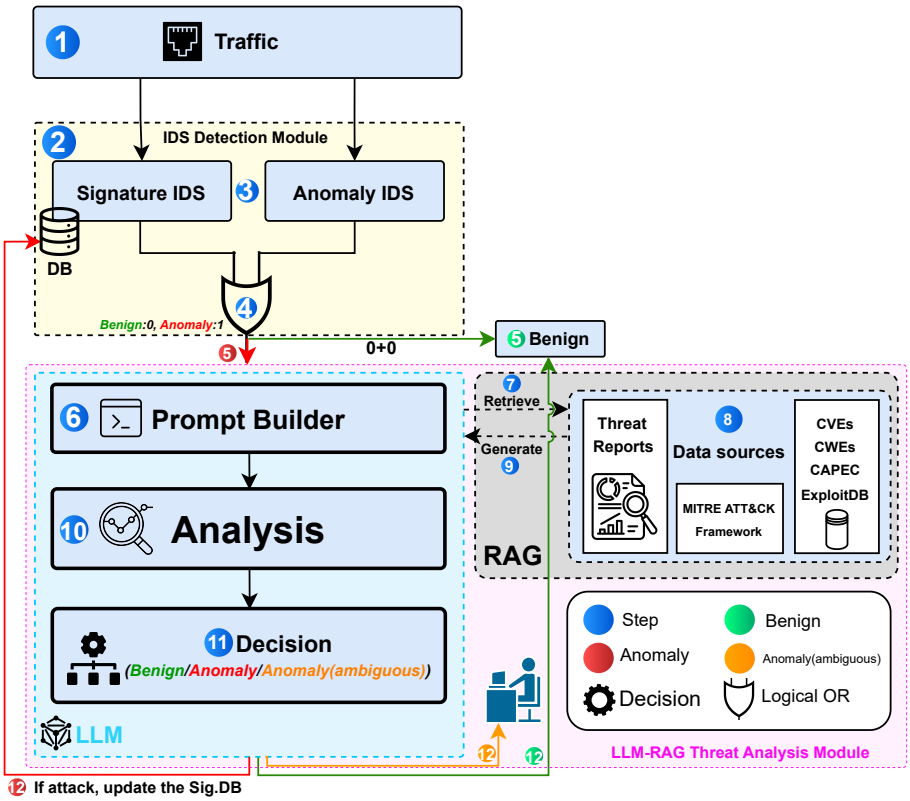

Fig. 1. The HyLLM-IDS framework

Anomaly-based IDS. (*step 3*) For the anomaly-based IDS, we employ an *unsupervised machine learning model* suitable for zero-day attack detection, where no prior labels exist. Possible machine learning model candidates include Isolation Forest (IF) [29] and Local Outlier Factor (LOF) [30], chosen for their effectiveness in high-dimensional anomaly detection.

3.2 LLM Analysis Module

Following a logical OR decision based on the outputs of both IDSs (*step 4*), and once the IDS module flags traffic as anomalous (*step 5 red*), a structured prompt is constructed using the prompt builder (*step 6*) as input to the LLM for further contextual analysis. The final prompt is submitted to the LLM (after utilizing either zero-shot learning or few-shot learning with Chain-of-Thought prompting), which analyzes (*step 10*) and classifies the traffic as one of three categories: Benign, Anomaly, or Anomaly (ambiguous) (*step 11*).

Anomaly (ambiguous) is returned as a default value when the LLM expresses low confidence in its classification. These cases are flagged for the security administrator for further analysis (*step 11 orange*). If the LLM confirms the presence of an anomaly, a detection rule is automatically generated and pushed to be added to the signature-based IDS database (*step 12 red*).

3.3 Retrieval-Augmented Generation (RAG)

As mentioned in the LLMs analysis module, while they offer powerful reasoning capabilities, they suffer from limitations such as hallucinations of incorrect facts, a static knowledge base, and are trained on public-domain data [31]. To address this, our framework integrates an RAG module (*steps 7-9*) that provides a semantic retrieval from real-time factual data sources about the suspicious traffic to augment the LLMs component.

In addition, RAG helps in reducing LLMs' hallucinations, which is one of the key challenges in LLMs. Our current RAG configuration queries three major sources:

1. Organised threat intelligence reports (e.g., ENISA, Mandiant, SANS, CrowdStrike, and CISA).
2. MITRE ATT&CK as a knowledge base of adversary tactics and techniques based on real-world observations.
3. Structured vulnerability databases (e.g., CVEs, CWEs, CAPEC, ExploitDB).

Further data sources can be integrated in the future, depending on changes in the scope of the targeted domain.
The overall HyLLM-IDS decision logic is summarised in Algorithm 1.

Algorithm 1 HyLLM-IDS Decision Logic (Parallel Dual Detection)

Input: Traffic T
Output: Classification $\in$ {**Benign**, **Anomaly**, **Anomaly (ambiguous)**}

$S \leftarrow$ SignatureBasedIDS(T) $A \leftarrow$ AnomalyBasedIDS(T)
if *S indicates anomaly* ***or*** *A indicates anomaly* **then**
  $L \leftarrow$ LLMAnalysis(T)
  if *L indicates anomaly* **then**
    UpdateSignatureDatabase(T) **return Anomaly**
  else if *L is ambiguous* **then**
    return Anomaly (ambiguous) ; `// Flag for human review`
  else
    return Benign
else
  return Benign

4 Discussion

The integration of the LLM-RAG threat analysis module is expected to significantly improve the recall and accuracy of intrusion detection by combining the contextual reasoning power of large language models with up-to-date information access by RAG to updated threat intelligence. This synergy enables more accurate classification of complex or ambiguous threats, particularly in dynamic cyber-physical environments. While the HyLLM-IDS framework offers notable advantages, its deployment introduces several implementation challenges. For example, the increased latency due to LLM inference and RAG retrieval overhead, the risk of misclassifications due to the reliance on retrieval quality that may access low-quality or incomplete knowledge bases, LLM vulnerabilities like prompt-based attacks, and the scarcity of high-quality labelled datasets for fine-tuning in domains like CPS applications. These challenges, along with proposed mitigation strategies, are summarized in Table 1 below.

Table 1. Challenges and Possible Mitigation Strategies

Challenge	Description	Possible Mitigation Strategy
Latency	Delay due to prompt building and analysis	Knowledge distillation, and Key-Value caching
Dependency on Dual IDS	LLM may not analyse traffic if both IDSs classify it as benign	Sample benign traffic periodically for LLM analysis
RAG Quality	Incomplete or poor retrieval can misclassify events	Regularly update and curate the RAG knowledge base
LLM Vulnerabilities	Susceptible to jailbreaks and prompt injection	Use input sanitisation and adversarial defence techniques
Domain Fine-tuning	Lacks quality labelled CPS-specific data	Use semi-supervised learning and data augmentation

5 Conclusion and Future Work

In this work, we introduced HyLLM-IDS, a hybrid intrusion detection system that integrates conventional signature-based and anomaly-based detection with the semantic reasoning capabilities of Large Language Models and Retrieval-Augmented Generation. The integration of a RAG module allows HyLLM-IDS to dynamically access up-to-date threat-related information, improving its responsiveness to emerging anomalies. Our framework aims to enhance both recall and precision in anomaly detection, while maintaining a balance between adaptability and accuracy, particularly in complex, evolving environments. In addition to the previously mentioned challenges and possible mitigation strategies in Table 1, future work will focus on the following key directions:

– **Foundation LLMs vs. Domain-Specific LLMs:** Investigate and compare the classification performance between auto-regressive models such as

Qwen3, *Gemma3*, or *Deepseek R1*, which rank highly on *huggingface.co* [32] and *lm-arena.ai* [33], both in their base and fine-tuned forms, with auto-encoding, domain-specific, security-pretrained models such as *SecBERT* and *SecureBERT*.
- **Benchmarking and Evaluation:** Utilize standard cybersecurity and NLP benchmarks to assess the overall performance across various complex attack scenarios.
- **Adaptive Sampling:** Investigate an intelligent traffic sampling to detect complex threats in benign-classified traffic.

Overall, HyLLM-IDS represents a promising step toward adaptive, intelligent, and context-aware intrusion detection systems for cyber-physical systems, laying a robust foundation for future developments of AI-driven cybersecurity solutions.

References

1. Vikas Khullar, Vrajesh Sharma, Mohit Angurala, and Nipun Chhabra, editors. *Artificial Intelligence and Society 5.0: Issues, Opportunities, and Challenges*. Chapman & Hall / CRC Press, an imprint of Taylor & Francis Group, LLC, Boca Raton, FL and Abingdon, Oxon, 2024. Available at: https://doi.org/10.1201/9781003397052.
2. Charalampos Patrikakis and Kincho Law. Society 5.0: Human centric, decentralized, and hyperautomated. *IT Professional*, 24(3):16–17, 2022. Available at: https://doi.org/10.1109/MITP.2022.3177281.
3. Amit Tyagi, R. Priya, Anand Mishra, and Balamurugan .G. *Industry 5.0: Potentials, Issues, Opportunities, and Challenges for Society 5.0*, pages 409–432. 11 2023. Available at: https://doi.org/10.1002/9781394213726.ch17.
4. Amirkhosro Vosughi, Ali Tamimi, Alexandra Beatrice King, Subir Majumder, and Anurag K. Srivastava. Cyber–physical vulnerability and resiliency analysis for DER integration: A review, challenges and research needs. *Renewable and Sustainable Energy Reviews*, 168(C), 2022. Available at: https://ideas.repec.org/a/eee/rensus/v168y2022ics1364032122006785.html.
5. Mamdouh Muhammad, Abdullah S. Alshra'a, and Reinhard German. Survey of cybersecurity in smart grids protocols and datasets. *Procedia Computer Science*, 241:365–372, 2024. 14th International Conference on Sustainable Energy Information Technology. Available at: https://doi.org/10.1016/j.procs.2024.08.049.
6. Bundesamt für Sicherheit in der Informationstechnik. IT-Sicherheitsgesetz 2.0 (IT-SiG 2.0). https://www.bsi.bund.de/DE/Das-BSI/Auftrag/Gesetze-und-Verordnungen/IT-SiG/2-0/it_sig-2-0_node.html, 2025. Accessed: 2025-04-25.
7. Bundesamt für Sicherheit in der Informationstechnik. BSI-Gesetz (BSIG) – Federal Office for Information Security Act, 2025. Available at: https://www.bsi.bund.de/EN/Das-BSI/Auftrag/Gesetze-und-Verordnungen/BSI-Gesetz/bsi-gesetz.html (Accessed: 2025-04-23).
8. Bundesamt für Sicherheit in der Informationstechnik. FAQ: Systeme zur Angriffserkennung (SzA). https://www.bsi.bund.de/EN/Themen/Regulierte-Wirtschaft/Kritische-Infrastrukturen/KRITIS-FAQ/FAQ-Systeme-Angriffserkennung/faq-systeme-angriffserkennung_node.html, 2025. Accessed: 2025-05-03.

9. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689, June 2024.
10. Future of Life Institute. Annex III – High-Risk AI Systems According to Article 6(2). https://artificialintelligenceact.eu/annex/3/, 2025. Accessed: 2025-04-25.
11. Future of Life Institute. Article 15 – Accuracy, Robustness and Cybersecurity. https://artificialintelligenceact.eu/article/15/, 2025. Accessed: 2025-04-25.
12. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. https://arxiv.org/abs/2303.18223, 2025.
13. Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. https://arxiv.org/abs/2212.10403, 2023.
14. Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. Available at: http://dx.doi.org/10.1145/3703155.
15. Lingjiao Chen, Matei Zaharia, and James Zou. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2), mar 12 2024. Available at: https://hdsr.mitpress.mit.edu/pub/y95zitmz.
16. Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. https://doi.org/10.36227/techrxiv.23589741.v1, 07 2023.
17. Nurul Fauzi, Fazmah Yulianto, and Hilal Nuha. The effectiveness of anomaly-based intrusion detection systems in handling zero-day attacks using adaboost, j48, and random forest methods. pages 57–62, 10 2023. Available at: https://doi.org/10.1109/APWiMob59963.2023.10365642.
18. Nour Moustafa, Benjamin Turnbull, and Kim-Kwang Raymond Choo. An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things. *IEEE Internet of Things Journal*, 6(3):4815–4830, 2019. Available at: https://doi.org/10.1109/JIOT.2018.2871719.
19. Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho. Flow-based benchmark data sets for intrusion detection. In *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS)*, pages 361–369. ACPI, 2017. Available at: https://www.researchgate.net/publication/317271077.
20. Canadian Institute for Cybersecurity, University of New Brunswick. CIC Datasets. https://www.unb.ca/cic/datasets/, 2025. Accessed: 2025-04-28.
21. Hoang-Cong-Thanh Nguyen, Xuan-Ha Nguyen, and Kim-Hung Le. An automated benchmarking framework for anomaly-based intrusion detection systems. *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2024. Available at: https://api.semanticscholar.org/CorpusID:272574707.

22. Moorthy Agoramoorthy, Ahamed Ali, D. Sujatha, Michael F, and Guruvugari Ramesh. An analysis of signature-based components in hybrid intrusion detection systems. pages 1–5, 12 2023. Available at: https://doi.org/10.1109/ICCEBS58601.2023.10449209.
23. Fazalur Rehman, Farhan Mushtaq, and Hafsah Zaman. A host-based intrusion detection: Using signature-based and ai-driven anomaly detection for enhanced cybersecurity*. In *2024 4th International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pages 1–7, 2024. Available at: https://doi.org/10.1109/ICoDT262145.2024.10740248.
24. Sidahmed Benabderrahmane, Petko Valtchev, James Cheney, and Talal Rahwan. Apt-llm: Embedding-based anomaly detection of cyber advanced persistent threats using large language models. https://arxiv.org/abs/2502.09385, 2025.
25. Rikhiya Ghosh, Hans-Martin von Stockhausen, Martin Schmitt, George Marica Vasile, Sanjeev Kumar Karn, and Oladimeji Farri. Cve-llm : Ontology-assisted automatic vulnerability evaluation using large language models. https://arxiv.org/abs/2502.15932, 2025.
26. Aoxiao Zhong, Dengyao Mo, Guiyang Liu, Jinbu Liu, Qingda Lu, Qi Zhou, Jiesheng Wu, Quanzheng Li, and Qingsong Wen. Logparser-llm: Advancing efficient log parsing with large language models. https://arxiv.org/abs/2408.13727, 2024.
27. Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. Audit-llm: Multi-agent collaboration for log-based insider threat detection. https://arxiv.org/abs/2408.08902, 2024.
28. OISF – Open Information Security Foundation. Suricata Features. https://suricata.io/features/, 2025. Accessed: 2025-04-30.
29. Lu, Haowen. Evaluating the performance of svm, isolation forest, and dbscan for anomaly detection. *ITM Web Conf.*, 70:04012, 2025. Available at: https://doi.org/10.1051/itmconf/20257004012.
30. Arya Adesh, Shobha G, Jyoti Shetty, and Lili Xu. Local outlier factor for anomaly detection in hpcc systems. *J. Parallel Distrib. Comput.*, 192(C), oct 2024. Available at: https://doi.org/10.1016/j.jpdc.2024.104923.
31. Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. Rag in health care: A novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2(1):AIra2400380, 2025. Available at: https://ai.nejm.org/doi/full/10.1056/AIra2400380.
32. Hugging face models. https://huggingface.co/models. Accessed: 2025-05-02.
33. Lm arena: Benchmark and compare open llms. https://lmarena.ai/. Accessed: 2025-05-02.