Poster: Enhancing GNN Robustness for Network Intrusion Detection via Agent-based Analysis

Zhonghao Zhan
Department of Computing
Imperial College London

Huichi Zhou
Department of Computing
Imperial College London

Hamed Haddadi
Department of Computing
Imperial College London

***Abstract*—Graph Neural Networks (GNNs) show great promise for Network Intrusion Detection Systems (NIDS), particularly in IoT environments, but suffer performance degradation due to distribution drift and lack robustness against realistic adversarial attacks. Current robustness evaluations often rely on unrealistic synthetic perturbations and lack demonstrations on systematic analysis of different kinds of adversarial attack, which encompass both black-box and white-box scenarios. This work proposes a novel approach to enhance GNN robustness and generalization by employing Large Language Models (LLMs) in an agentic pipeline as simulated cybersecurity expert agents. These agents scrutinize graph structures derived from network flow data, identifying and potentially mitigating suspicious or adversarially perturbed elements before GNN processing. Our experiments, using a framework designed for realistic evaluation and testing with a variety of adversarial attacks including a dataset collected from physical testbed experiments, demonstrate that integrating LLM analysis can significantly improve the resilience of GNN-based NIDS against challenges, showcasing the potential of LLM agent as a complementary layer in intrusion detection architectures.**

***Index Terms*—Graph Neural Network, Large Language Model, Network Intrusion Detection, Adversarial Attack, Robustness**

I. Introduction

Graph Neural Networks (GNNs) offer significant potential for Network Intrusion Detection Systems (NIDS) by modeling complex network relationships. However, realizing their practical utility requires addressing critical robustness challenges often overlooked in standard evaluations. Our work tackles two primary issues revealed through rigorous preliminary analysis.

First, we address the challenge of generalization under distribution drift. Standard GNN evaluation often relies on training and testing within single, static datasets (e.g., a specific version of CIC-IDS or UNSW-NB15). This does not capture performance in real-world networks where traffic patterns constantly evolve [1]. To investigate this, we constructed a *unified dataset* by merging and standardizing flows from various canonical sources, including UNSW-NB15 [2], CIC-IDS2018 [3], and Bot-IoT [4]. Evaluating representative GNN models (such as EGraphSage [5], Anomal-E [6], and CAGN-GAT [7]) using this unified dataset revealed a notable performance degradation compared to their single-dataset benchmarks, empirically confirming their susceptibility to distribution drift, a fundamental robustness problem.

Second, beyond drift, we examine robustness against adversarial attacks under realistic conditions. While GNNs might claim robustness, evaluations often use synthetic attacks, particularly white-box methods, that grant attackers unrealistic capabilities [8]. To probe vulnerability more realistically, we simulated practical, problem-space black-box attacks (such as node injection) on our unified dataset. These experiments demonstrated *further* significant performance degradation in the evaluated GNN models, exposing their vulnerability even to attacks feasible for real adversaries. This finding underscores the inadequacy of current robustness claims and highlights the need for both more realistic adversarial testing scenarios and effective mitigation strategies.

The observed GNN fragility against both drift and realistic attacks motivates our core proposal: leveraging LLMs as simulated cybersecurity experts. Given that human experts can often identify sophisticated attacks or anomalies missed by automated systems [9], [10], we explore using LLM agents to perform preemptive analysis on network flow graph data. By scrutinizing graph structures and flow patterns, the LLM agent aims to filter threats or alert operators before they compromise the GNN. This poster presents our investigation into this LLM-agent approach, detailing its integration and presenting experimental results that demonstrate its effectiveness in enhancing GNN resilience against simulated realistic attacks, offering a promising path towards more robust NIDS deployment.

II. Related Work

Current evaluations of GNN-based NIDS face limitations hindering realistic assessment. Models are often evaluated on single static datasets, ignoring performance degradation due to distribution drift [1]. Furthermore, adversarial robustness assessments frequently rely on synthetic perturbations generated without considering network domain constraints [8], potentially rendering attacks impractical [8]. Much GNN robustness research uses generic graph benchmarks rather than NIDS-specific structures [11], and realistic structural attacks remain underexplored [8]. Existing GNN robustness benchmarks may also lack sufficient integration of drift or domain-specific realism [12]. Ideas leveraging LLMs for robust evaluation or enhancing GNNs are emerging [13], [14], but applying them as expert simulators for NIDS remains unexplored. This evaluation gap leads to an overestimation of GNN resilience.

III. Methodology

Our approach integrates LLM analysis into the GNN-based NIDS pipeline.

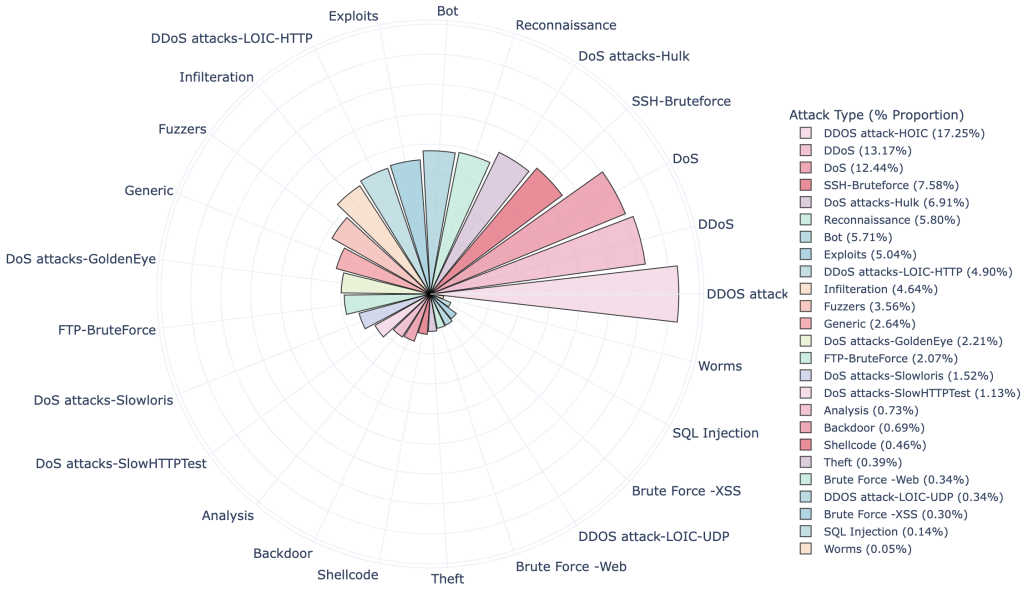

Fig. 1: The intrusion types of unified dataset for distribution drift test

TABLE I: Comparison of GNN Performance Across Datasets

Model	Dataset	Accuracy	F1
CAGN	Unified	0.851	0.823
	UNSW-NB15	0.995*	0.918*
	CICIDS2017	0.975*	0.881*
AnomalE	Unified	0.861	0.875
	UNSW-NB15	0.987*	0.924*
	CICIDS2018	0.971*	0.924*
E-GraphSage	Unified	0.675	0.793
	NF-ToN-IoT	0.672*	0.810*
	NF-Bot-IoT	0.782*	0.630*

* Results reported in the original paper of the models in the table. The 'Unified' results reflect performance on the unified dataset used in this study. AnomalE F1 for Combined are weighted averages.

A. Dataset and GNN Models

We utilized a unified dataset constructed by merging and standardizing multiple NetFlow-based intrusion detection datasets (NF-BoT-IoT [4], NF-CSE-CIC-IDS2018 [3], NF-UNSW-NB15 [2]) to simulate distribution drift. This allows for evaluating GNN generalization across diverse network scenarios. We evaluated several GNN architectures commonly used in NIDS, including E-GraphSAGE [5], Anomal-E [6], and CAGN-GAT [7]. Network flows were transformed into IP-centric communication graphs where nodes represent IPs and edges represent flows with associated features.

B. LLM Agent Integration

We designed an LLM-based mitigation strategy where LLM agents analyze elements of the network graph before GNN processing. This involves:

- **Task Definition:** The LLM's task is to assess the relevance or potential maliciousness of graph components, such as injected nodes simulating certain attacks or suspicious edges/flows (conceptualized in Figure 2).
- **Prompting Strategy:** LLMs are prompted with textual information representing graph elements (e.g., descriptions of node interactions or flow characteristics) and asked to provide an analysis and relevance score.
- **Workflow:** The LLM acts as a filter or expert advisor. Based on the LLM's assessment, suspicious graph elements could be flagged, removed, or weighted differently before being fed into the GNN model. We specifically tested this against node injection attacks where the LLM attempts to identify the injected malicious nodes.

C. Evaluation Protocol

We evaluated the GNN models under several conditions:

1) Baseline performance on standardized datasets.
2) Performance under simulated distribution drift using the unified dataset.
3) Robustness against synthetic attacks (PGD feature attacks, Edge Removal, Node Injection).
4) Performance of the GNN with LLM-based mitigation applied, particularly focusing on the recovery from node injection attacks.

Metrics included Accuracy, F1-Score, Precision, Recall, and AUC, with a focus on F1-Score due to class imbalance inherent in NIDS datasets.

IV. Results and Discussion

Our experiments confirm the susceptibility of standard GNN models to distribution drift and synthetic attacks. Node injection, in particular, is observed to degrade accuracy and F1-scores across the evaluated models.

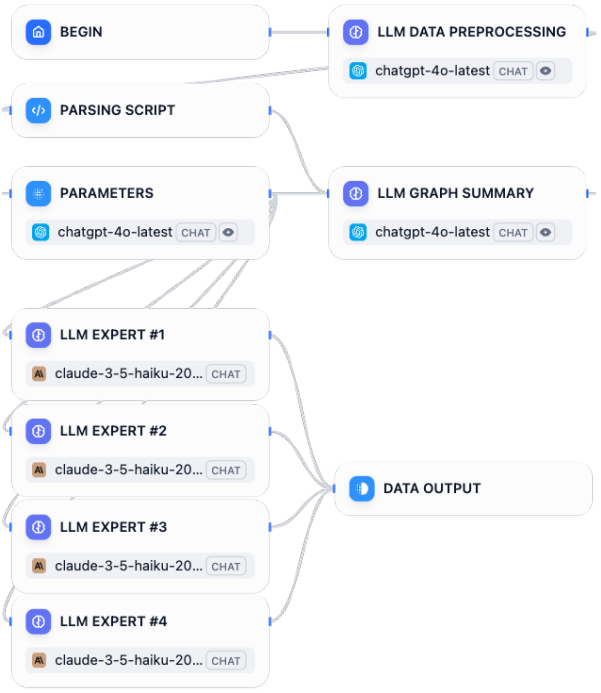

Fig. 2: Design diagram of the LLM Mitigation pipeline

The key finding highlighted in this poster is the effectiveness of the LLM mitigation strategy. LLMs (specifically tested models like GPT-4o and LLaMA 4) demonstrated a strong capability to identify artificially injected malicious nodes within the graph structure derived from netflow data, as shown in Table II. For instance, in scenarios with 20% node injection (200 injected nodes into a base graph of 1000), leading LLMs correctly flagged a high percentage of these malicious nodes.

By filtering or identifying these malicious nodes based on LLM analysis, the downstream GNN classification performance improved significantly compared to the attacked scenario without LLM mitigation (Table II). While tested against synthetic node injection, the LLM's ability to reason about network interactions based on provided data hints at potential benefits for handling novel, zero-day attacks or complex drift patterns that manifest as unusual graph structures.

TABLE II: Performance of LLMs in graph mitigation. CAGN-GAT Fusion [7] is being tested. CF = Correctly Flagged, IF = Incorrectly Flagged

Model	Accuracy	F1	LLM Recall	Nodes	CF	IF
Clean	0.842	0.821	–	1000	–	–
Claude 3.5 Haiku	0.777	0.673	0.698	1036	0	164
Claude 3.7 Sonnet	0.774	0.728	0.741	1107	35	**58**
LLaMA 4 Maverick 17B	**0.859**	0.834	**0.758**	931	**200**	69
GPT-4o	0.857	**0.838**	0.774	945	197	**58**

V. Conclusion

Standard GNN evaluations often overlook performance degradation from distribution drift and employ unrealistic adversarial attacks. Our analysis using a unified dataset confirms GNN susceptibility to drift, with further degradation under simulated realistic attacks. This work demonstrates that integrating *individual* LLM as expert analyzers enhances GNN resilience against attacks like node injection, showcasing the potential of hybrid GNN-LLM systems.

Future work will focus on developing an LLM agentic pipeline. This pipeline aims to leverage different models strategically: for instance, employing GPT-4o for complex, high-level graph analysis to generate context that guides more cost-efficient models, such as Claude 3.5 Haiku, for scaled processing. The goal is to create a cost-effective workflow that maintains strong detection performance against network intrusions and adversarial attacks. Furthermore, we plan to validate these findings against diverse, realistic attack scenarios within a physical IoT testbed, utilizing real-world tools (e.g., Kali Linux) and incorporating physical device manipulations. Optimizing the overall pipeline efficiency for potential real-time NIDS deployment remains a crucial final objective.

References

[1] G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, and L. Cavallaro, "Insomnia: Towards concept-drift robustness in network intrusion detection," in *Proceedings of the 14th ACM workshop on artificial intelligence and security*, 2021, pp. 111–122.

[2] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[3] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.

[4] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[5] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, and M. Portmann, "E-graphsage: A graph neural network based intrusion detection system for iot," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–9.

[6] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, "Anomal-e: A self-supervised network intrusion detection system based on graph neural networks," *Knowledge-based systems*, vol. 258, p. 110030, 2022.

[7] M. A. Jahin, S. Soudeep, M. Mridha, R. Kabir, M. R. Islam, and Y. Watanobe, "Cagn-gat fusion: A hybrid contrastive attentive graph neural network for network intrusion detection," *arXiv preprint arXiv:2503.00961*, 2025.

[8] J. Ma, S. Ding, and Q. Mei, "Towards more practical adversarial attacks on graph neural networks," *Advances in neural information processing systems*, vol. 33, pp. 4756–4766, 2020.

[9] W. Ding, H. Hu, and L. Cheng, "Iotsafe: Enforcing safety and security policy withreal iot physical interaction discovery," in *Network and Distributed System Security Symposium*, 2021.

[10] S. Siboni, V. Sachidananda, A. Shabtai, and Y. Elovici, "Security testbed for the internet of things," *arXiv preprint arXiv:1610.05971*, 2016.

[11] C. Wang, P. Zheng, J. Gui, C. Hua, and W. U. Hassan, "Are we there yet? unraveling the state-of-the-art graph network intrusion detection systems," *arXiv preprint arXiv:2503.20281*, 2025.

[12] Q. Zheng, X. Zou, Y. Dong, Y. Cen, D. Yin, J. Xu, Y. Yang, and J. Tang, "Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning," *arXiv preprint arXiv:2111.04314*, 2021.

[13] H. Zhou, K.-H. Lee, Z. Zhan, Y. Chen, Z. Li, Z. Wang, H. Haddadi, and E. Yilmaz, "Trustrag: Enhancing robustness and trustworthiness in rag," *arXiv preprint arXiv:2501.00879*, 2025.

[14] Z. Zhang, X. Wang, H. Zhou, Y. Yu, M. Zhang, C. Yang, and C. Shi, "Can large language models improve the adversarial robustness of graph neural networks?" *arXiv preprint arXiv:2408.08685*, 2024.