Towards Effective Identification of Attack Techniques in Cyber Threat Intelligence Reports using Large Language Models

Hoang Cuong Nguyen
Swinburne University of Technology
Melbourne, Australia

Shahroz Tariq*
CSIRO's Data61
Sydney, Australia

Mohan Baruwal Chhetri
CSIRO's Data61
Melbourne, Australia

Bao Quoc Vo
Swinburne University of Technology
Melbourne, Australia

Abstract

This work evaluates the performance of Cyber Threat Intelligence (CTI) extraction methods in identifying attack techniques from threat reports available on the web using the MITRE ATT&CK framework. We analyse four configurations utilising state-of-the-art tools, including the Threat Report ATT&CK Mapper (TRAM) and open-source Large Language Models (LLMs) such as Llama2. Our findings reveal significant challenges, including class imbalance, overfitting, and domain-specific complexity, which impede accurate technique extraction. To mitigate these issues, we propose a novel two-step pipeline: first, an LLM summarises the reports, and second, a retrained SciBERT model processes a rebalanced dataset augmented with LLM-generated data. This approach achieves an improvement in F1-scores compared to baseline models, with several attack techniques surpassing an F1-score of 0.90. Our contributions enhance the efficiency of web-based CTI systems and support collaborative cybersecurity operations in an interconnected digital landscape, paving the way for future research on integrating human-AI collaboration platforms.

Keywords

Cyber Threat Intelligence, LLMs, Text Summarisation, Cybersecurity, MITRE ATT&CK Techniques, Cyber Reports

1 Introduction

In today's rapidly evolving digital landscape, cybersecurity has emerged as a critical concern for organisations worldwide. Security Operations Centres (SOCs) play a pivotal role in defending against the increasing sophistication of cyber threats by leveraging advanced technologies, such as artificial intelligence and machine learning [8]. These technologies enhance the capacity to detect, analyse, and respond to threats in real time, thereby improving the resilience of digital infrastructures. Moreover, the integration of human-AI teaming and collaboration is gaining traction as a strategy to enhance the efficiency and effectiveness in different domains [7, 13, 16] including cybersecurity operations[1, 5, 19, 20]. By combining the analytical strengths of human experts with the rapid processing capabilities of AI, SOCs can more effectively manage the vast and complex data streams they encounter daily. Cybersecurity analysts often rely on Cyber Threat Intelligence (CTI) reports to remain informed about the ever-evolving threat landscape.

CTI reports are comprehensive documents that provide valuable insights into current and emerging cyber threats faced by organisations. Typically produced by cybersecurity analysts or specialised agencies, these reports aid businesses and government entities in understanding the threat landscape and adopting proactive measures to safeguard their digital assets [15]. Key components of a CTI report include (i) detailed descriptions of various cyber threats, such as malware, ransomware, phishing attacks, and advanced persistent threats (APTs); (ii) profiles of threat actors, including their motives, tactics, techniques, and procedures (TTPs); (iii) Indicators of Compromise (IOCs) like IP addresses, malware hashes, or domain names that signal potential breaches; and (iv) recommended mitigation strategies to counteract these threats.

However, the manual analysis of CTI reports poses significant challenges due to their often unstructured and verbose nature. Such reports can extend over dozens of pages, making it arduous for SOC analysts to swiftly extract critical information [17]. This inefficiency contributes to the broader issue of alert fatigue, with studies indicating that up to 70% of SOC analysts feel overwhelmed by the volume of alerts, leading 43% to disable alerts as a coping mechanism [18, 22]. Given that modern cybersecurity operations depend heavily on real-time, web-based collaboration and decision-making, addressing these inefficiencies is vital for sustaining effective threat defence.

To alleviate these challenges, automated CTI extraction methods have been developed, facilitating the identification of IOCs and TTPs from extensive web-sourced reports [4, 9]. Despite advancements in AI and Natural Language Processing (NLP), several hurdles persist in automating CTI analysis: **(i) Domain Complexity**: CTI reports often contain specialised terminology distinct from standard English, hindering accurate extraction by generic NLP tools; **(ii) Verbosity**: The relevant information about cyber-attacks is often buried within lengthy documents. For instance, a 42-page report [6] may only dedicate a few paragraphs to the actual attack details; and **(iii) Relationship Extraction**: Accurately capturing the relationships between entities, such as attackers, tools, and victims, is essential for understanding TTPs, yet current NLP systems struggle with this intricate task [10].

This study aims to address these challenges by exploring innovative approaches that enhance the automated extraction and utilisation of CTI reports, ultimately empowering SOCs to make more informed and timely decisions in the face of evolving cyber threats. Given these challenges, this research explores the following research questions:

[*] Corresponding Author. Email: shahroz.tariq@data61.csiro.au

Accepted for publication at The Web Conference 2025 (WWW '25).

- **RQ1:** How effective are standalone vanilla large language models (LLMs) in CTI extraction?
- **RQ2:** Can LLM-based augmentation improve the performance of automated CTI extraction methods?

In answering these research questions, we make the following contributions:

(1) **Comprehensive Evaluation**: We evaluate the CTI extraction method using four configurations, as shown in Figure 1. Our evaluation highlights the strengths and weaknesses of each configuration, with the best-performing baseline (i.e., TRAM with original SciBert), achieving an F1-score of just over 0.4 due to overfitting and class imbalances.

(2) **Novel Extraction Pipeline**: Inspired by recent developments in LLMs, we propose a two-step pipeline: *(i) CTI Report Summarisation using GPT-3.5*: Reducing report verbosity to focus on key threat information; and *(ii) Retrained SciBERT Model*: Using a SciBERT model [2] trained on a rebalanced dataset to address class imbalance and improve classification accuracy, resulting in an F1-score of over 0.90 in identification of several attack techniques.

The remainder of this paper is organised as follows: Section 2 reviews the current progress and limitations in CTI extraction. Section 3 details the evaluation methodology and proposed approach. Section 4 presents the evaluation results. Section 5 concludes this work and outlines the limitations and future research directions[1].

2 Background and Related Works

This section provides an overview of CTI, its extraction and sharing processes, and the current limitations of existing CTI extraction methods, highlighting the need for improved methodologies.

2.1 Cyber Threat Intelligence

Cyber Threat Intelligence (CTI) serves as a preventative defence mechanism against cyber-attacks. According to the National Institute of Standards and Technology (NIST), CTI is defined as “threat information that has been aggregated, transformed, analysed, interpreted, or enriched to provide the necessary context for decision-making processes” [11].

The primary criterion for evaluating CTI is its actionability. According to Pawlinski et al. [14], actionable CTI must possess the following characteristics: *(i) Relevance*: Applicable to the area of responsibility for recipients; *(ii) Timeliness*: Information must be recent enough to be effective; delays can render CTI obsolete; *(iii) Accuracy*: Information should be verified and error-free; *(iv) Completeness*: Sufficient context to understand past cyber-attacks; and *(v) Ingestibility*: Shared in a format that can be processed by recipient systems.

2.2 CTI Extraction and Sharing

Modern CTI extraction methods typically follow a standardised pipeline comprising the following steps: *(i) Identifying Sources*: Selecting relevant threat reports for analysis; *(ii) Report Crawling*: Automatically gathering reports from various repositories; *(iii) Text Processing and Labelling*: Extracting and annotating relevant entities, such as Indicators of Compromise (IOCs) and Tactics, Techniques, and Procedures (TTPs); *(iv) Text Summarisation*: Reducing verbosity while retaining key information using learning-based approaches; and *(v) Processing Outputs*: Converting extracted data into formats suitable for downstream applications, such as knowledge graphs. Methods such as AttackG [9] and TRAM [4] have been developed to automate CTI extraction. These methods leverage NLP techniques to identify attack techniques based on the MITRE ATT&CK framework.

Current Limitations of CTI Extraction Methods. Despite advancements, CTI extraction methods face significant challenges that hinder their effectiveness: *(i) Domain Complexity*: CTI reports contain cybersecurity-specific terminology that differs from standard English, making it difficult for general NLP models to process accurately. *(ii) Verbosity*: Many CTI reports are lengthy, with only a small portion containing actionable threat information. For example, a 42-page report may contain only a few sentences detailing the actual attack [6]. *(iii) Relationship Extraction*: Extracting relationships between entities (e.g., attackers, tools, victims) is essential but remains challenging for existing NLP systems [10]. *(iv) Class Imbalance*: Many extraction methods suffer from class imbalance, where certain techniques are overrepresented while others are underrepresented. *(v) Replication Inconsistency*: Performance claims in research papers often fail to replicate on different datasets due to varying conditions and datasets.

These challenges contribute to low precision, recall, and F1-scores, highlighting the need for more effective CTI extraction methodologies. Our work attempts to address some of the limitations of existing CTI extraction methods through a comprehensive evaluation of multiple CTI extraction methods and the introduction of a novel extraction pipeline.

3 Methodology

Our evaluation methodology consists of five key components: extraction method selection, dataset preparation, experimental design, and evaluation metrics.

3.1 Extraction Method Selection

To answer **RQ1**, we selected three variants of vanilla Llama2 [21], i.e., 7B, 13B, and 70B, under zero-shot prompting. To answer **RQ2**, we selected TRAM, a SciBERT-based method designed to classify sentences into the 50 most prevalent techniques within the MITRE ATT&CK framework [4], as a base model and used different configurations with and without the LLMs-based augmentations for evaluation.

3.2 Ground Truth Datasets

Two annotated datasets were used as ground truth for evaluation:

(1) ***Adversary Emulation Library (AEL)***: This dataset comprises concise reports on attack campaigns (e.g., APT29, Carbanak, FIN6) annotated with MITRE ATT&CK technique IDs [3].

(2) ***Attack-Technique-Dataset (ATD)***: This dataset contains longer reports (e.g., OceanLotus, Sowbug, MuddyWater) annotated with detailed technique information [12].

[1] Our code is available here: https://github.com/hoangcuongnguyen2001/SciBERT-for-Technique-Classification

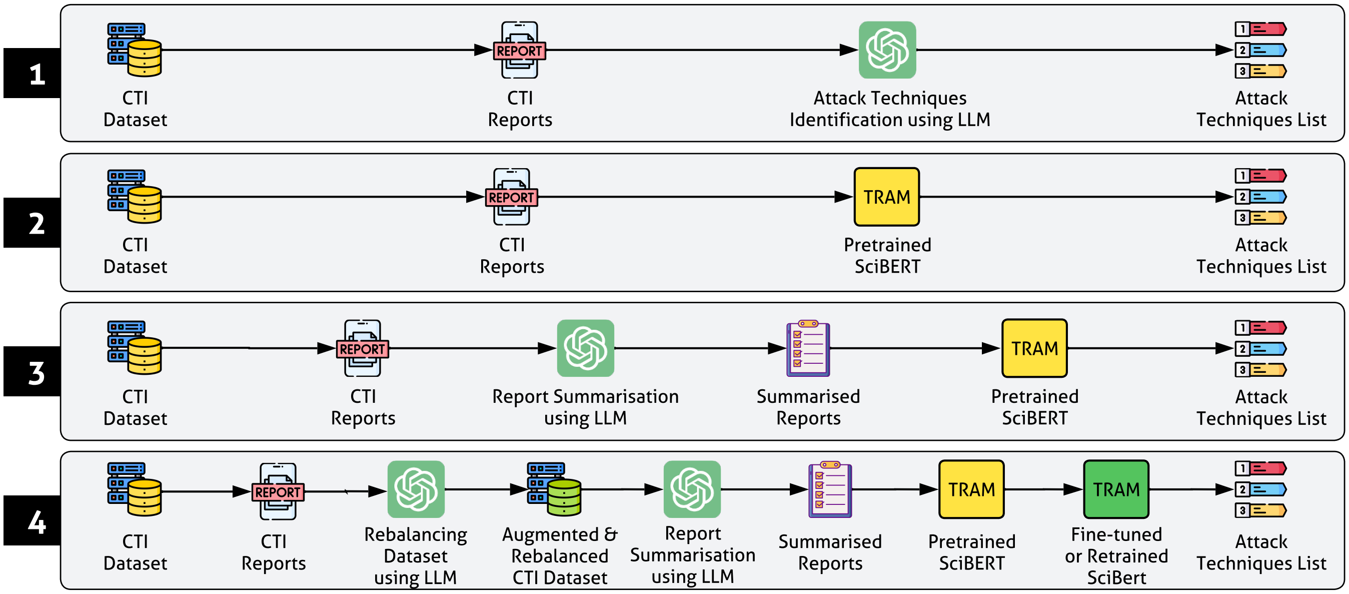

Figure 1: Our evaluation methodology uses four configurations.

Reports were preprocessed to remove technique IDs, hyperlinks, and extraneous content to ensure unbiased evaluation. Techniques outside the top 50 most prevalent techniques in the MITRE ATT&CK framework were excluded to align with TRAM's training data.

3.3 Experimental Design

As shown in Figure 1, we use four experimental settings to assess and compare the CTI extraction methodologies:

(1) ***Standalone LLM for CTI Extraction***: The zero-shot prompting capabilities of open-source LLMs (Llama2) were evaluated. Precision, recall, and F1-score were computed, along with counts of true positives, false positives, and false negatives.

(2) ***Original TRAM Configuration***: TRAM's performance was evaluated using a pre-trained SciBERT model with confidence thresholds of 25% and 80%. We will denote it as Original SciBERT.

(3) ***TRAM with LLM-based Summarisation***: This configuration uses summarised CTI reports generated by an LLM (GPT-3.5), followed by SciBERT classification at confidence levels of 25% and 75%. We will denote it as aCTIon. *Note: This selection is based on best performance settings.*

(4) ***TRAM with LLM-based Summarisation, Rebalancing and Retraining***: In this configuration, we are Augmenting underrepresented techniques using GPT-3.5 and downsampling overrepresented techniques. CTI reports were summarised using GPT-3.5 to reduce verbosity and retain relevant content related to attack techniques. Then, summarised reports were processed by a SciBERT model retrained on the rebalanced dataset. We used three settings for training i.e., retraining, fine-tuning and retraining with 5-fold cross-validation, resulting in a total of 9 configurations for evaluation of TRAM.

Table 1: Comparison of false positives and false negatives between Llama2 models for AEL dataset. The '-' sign represents a false negative and the '+' sign represents false positives.

Report	Llama2-7B	Llama2-13B	Llama2-70B
APT29	-2 / +9	-2 / +10	0 / +9
Carbanak	-1 / +5	-5 / +8	-3 / +8
FIN6	-21 / +7	-24 / +8	-22 / +19
FIN7	-4 / +7	-3 / +14	-2 / +7
menuPass	-4 / +5	-4 / +11	-5 / +12
OilRig	-1 / +5	-2 / +5	-2 / +8

Evaluation Metrics. The performance of each extraction method was evaluated using standard classification metrics: precision, recall and F1-score. *Note: We only report the F1-score in this work due to space constraints.* Additional analysis included counts of true positives, false positives, and false negatives. For AttackG and LLMs, results were considered correct if the extracted technique names or IDs matched the ground truth. For TRAM, exact matches of both technique names and IDs were required.

4 Results

This section presents the evaluation results of standalone open-source LLMs for CTI extraction, followed by a comparative analysis of different TRAM configurations, highlighting the impact of model refinements and dataset rebalancing on performance.

4.1 Evaluation of Standalone Open-Source LLMs

To address RQ1, we evaluate the performance of LLama2 using six reports from the AEL dataset, each under 500 words. This ensured

Table 2: True positives for Llama2 models.

Report	Ground Truth	Llama2 7B	Llama2 13B	Llama2 70B
APT29	3	1	0	3
Carbanak	6	5	1	3
FIN6	24	3	0	2
FIN7	4	0	1	2
menuPass	5	1	1	0
OilRig	4	3	2	2

Table 3: Average performance of Llama2 models.

Method	Precision	Recall	F1-Score
Llama2-7B	**0.2403**	0.3736	**0.2733**
Llama2-13B	0.0911	0.2792	0.1199
Llama2-70B	0.1733	**0.4306**	0.2384

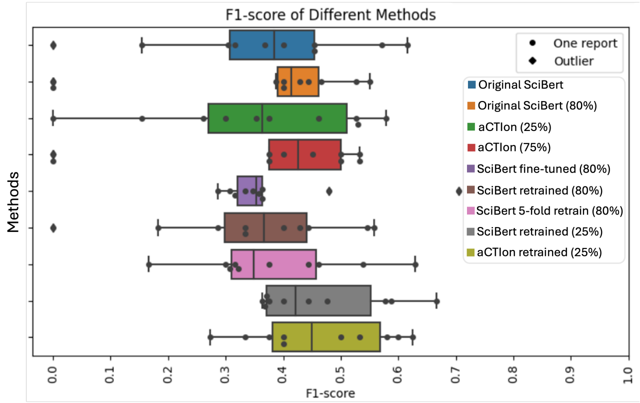

Figure 2: Performance of TRAM using different Configurations proposed in our work on ATD dataset.

manageable processing times and mitigated dependency explosion issues. Table 1 compares the number of false positives and false negatives for various versions of LLama2 (7B, 13B, and 70B). Notably, the larger 70B LLama2 model tended to produce more false positives in several instances. For example, in the case of FIN6, LLama2-7B, and LLama2-13B produced 7, and 8 false positives, respectively, whereas LLama2-70B produced 19 false positives, more than twice that of LLama2-13B. Future work should explore whether fine-tuning the LLama2 model could reduce the number of false positives and negatives.

Table 2 displays the number of true positives detected by each model and the overall precision, recall, and F1-score for each method are summarised in Table 3. All models performed poorly, detecting only a fraction of the true positives in many scenarios, as evident in Tables 2 and 3. Although the LLama2 70B model detected the most true positives, the overall performance of LLama2 7B was slightly better due to the higher number of false positives generated by LLama2 70B. This low performance highlights the complexity of the task and the need for specialised methods for CTI extraction. Our next research question aims to explore this very direction.

Table 4: Classification results for selected techniques using the retrained SciBERT model on the ATD dataset.

Technique ID	Precision	Recall	F1-Score	Sentences in Test Set
T1056.001	0.8000	0.8649	0.8312	37
T1057	0.7931	**1.0000**	0.8846	46
T1059.003	0.6848	0.7975	0.7368	79
T1070.004	0.9000	0.9474	**0.9231**	76
T1566.001	0.8679	0.9583	0.9109	48
T1570	**1.0000**	0.3125	0.4762	16

4.2 Comparison of TRAM Configurations

To address RQ2, we evaluated TRAM under different configurations and confidence levels (25% and 80%). The results, shown in Figure 2, highlight the impact of configuration changes. We can observe that first summarising the report with GPT-3.5 and then SciBERT classification results in slightly better performance than the default setting of SciBERT.

To address class imbalance and overfitting in TRAM, we retrained the SciBERT model on a rebalanced dataset. This approach resulted in a median F1-score increase of approximately seven percentage points compared to the baseline model, as shown in Figure 2. Classification results for a few selected techniques using the best-performing retrained SciBERT model are presented in Table 4. We observed that this model performs well on many of the top 50 most prevalent techniques in the MITRE ATT&CK framework, achieving an F1-score of up to 0.92.

5 Conclusion

This work evaluated state-of-the-art Cyber Threat Intelligence extraction methods, highlighting key challenges such as class imbalance, overfitting, and the complexity of cybersecurity texts. To address these issues, we proposed a novel pipeline that combines GPT-3.5 for report summarisation and a retrained SciBERT model to improve classification accuracy using rebalanced data augmented by an LLM. This approach resulted in a seven-percentage-point increase in the F1-score compared to the baseline model and achieved above 0.90 F1-score for several attack techniques. Despite these improvements, challenges remain in classifying underrepresented techniques and reducing false positives. Future work should focus on integrating human-AI collaboration to enhance extraction accuracy and exploring fine-tuned LLMs for more effective CTI analysis. These contributions lay the groundwork for more reliable automated CTI systems to support cybersecurity operations.

Acknowledgments

This work was supported by Commonwealth Scientific and Industrial Research Organisation (CSIRO) Collaborative Intelligence (CINTEL) Future Science Platform. The diagram have been designed using images from flaticon.com

References

[1] Mohan Baruwal Chhetri, Shahroz Tariq, Ronal Singh, Fatemeh Jalalvand, Cecile Paris, and Surya Nepal. 2024. Towards human-ai teaming to mitigate alert fatigue

in security operations centres. *ACM Transactions on Internet Technology* 24, 3 (2024), 1–22.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[3] Center for Threat-Informed Defense. 2024. Adversary Emulation Library. https://github.com/center-for-threat-informed-defense/adversary_emulation_library Accessed: 2024-06-12.

[4] Center for Threat-Informed Defense. 2024. TRAM: Threat Report ATT&CK Mapping. https://github.com/center-for-threat-informed-defense/tram Accessed: 2024-06-12.

[5] M Baruwal Chhetri, S Tariq, R Singh, F Jalalvand, C Paris, S Nepal, AA Vali, S Azizi, M Shojafar, S Saryazdi, et al. 2024. Internet Technology. *ACM Transactions on* 24, 3 (2024).

[6] ClearSky Cyber Security. 2016. *Operation DustySky*. Technical Report. https://www.clearskysec.com/wp-content/uploads/2016/01/Operation%20DustySky_TLP_WHITE.pdf Accessed: 2024-06-12.

[7] Jessica Irons, Patrick Cooper, Melanie McGrath, Shahroz Tariq, and Andreas Duenser. 2024. Towards a criteria-based approach to selecting human-AI interaction mode. *arXiv preprint arXiv:2411.07406* (2024).

[8] Fatemeh Jalalvand, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2024. Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods. *Comput. Surveys* 57, 2 (2024), 1–36.

[9] Zhenyuan Li, Jun Zeng, Yan Chen, and Zhenkai Liang. 2022. AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports. In *European Symposium on Research in Computer Security*. Springer, 589–609.

[10] Dongliang Mu, Alejandro Cuevas, Limin Yang, Hang Hu, Xinyu Xing, Bing Mao, and Gang Wang. 2018. Understanding the reproducibility of crowd-reported security vulnerabilities. In *27th USENIX Security Symposium (USENIX Security 18)*. 919–936.

[11] National Institute of Standards and Technology (NIST). 2024. Threat Intelligence - Glossary Term. https://csrc.nist.gov/glossary/term/threat_intelligence Accessed: 2024-06-12.

[12] NewBee119. 2024. Attack Technique Dataset. https://github.com/NewBee119/Attack-Technique-Dataset Accessed: 2024-06-12.

[13] Cécile Paris and Andrew Reeson. 2024. What's the Secret to Making Sure AI Does Not Steal Your Job? Work with It, Not Against It. In *The Conversation on Work*, Ian O. Williamson (Ed.). Johns Hopkins University Press, Baltimore, 177–181. First published in The Conversation on November 30th, 2021.

[14] Paweł Pawlinski, Przemylaw Jaroszewski, Piotr Kijewski, Lukasz Siewierski, Pawel Jacewicz, Przemyslaw Zielony, and Radoslaw Zuber. 2014. Actionable information for security incident response. *European Union Agency for Network and Information Security, Heraklion, Greece* (2014).

[15] Md Rayhanur Rahman, Rezvan Mahdavi Hezaveh, and Laurie Williams. 2023. What are the attackers doing now? Automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. *Comput. Surveys* 55, 12 (2023), 1–36.

[16] Emma Schleiger, Claire Mason, Claire Naughtin, Andrew Reeson, and Cecile Paris. 2024. Collaborative Intelligence: A scoping review of current applications. *Applied Artificial Intelligence* 38, 1 (2024), 2327890.

[17] BinHui Tang, JunFeng Wang, Zhongkun Yu, Bohan Chen, Wenhan Ge, Jian Yu, and TingTing Lu. 2022. Advanced Persistent Threat intelligent profiling technique: A survey. *Computers and Electrical Engineering* 103 (2022), 108261.

[18] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities. *ACM Comput. Surv.* 57, 9, Article 224 (April 2025), 38 pages. https://doi.org/10.1145/3723158

[19] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2024. A2C: A Modular Multi-stage Collaborative Decision Framework for Human-AI Teams. *arXiv preprint arXiv:2401.14432* (2024).

[20] Shahroz Tariq, Ronal Singh, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. Bridging Expertise Gaps: The Role of LLMs in Human-AI Collaboration for Cybersecurity. *arXiv* (2025).

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[22] Trend Micro. 2021. Cybersecurity Tool Sprawl Drives Plans to Outsource Detection and Response. https://newsroom.trendmicro.com/2021-10-12-Cybersecurity-Tool-Sprawl-Drives-Plans-to-Outsource-Detection-and-Response Accessed: 2024-06-12.