Extended Abstract: Assessing Language Models for Semantic Textual Similarity in Cybersecurity

Arian Soltani, DJeff Kanda Nkashama, Jordan Felicien Masakuna, Marc Frappier, Pierre-Martin Tardif, and Froduald Kabanza

GRIC, Université de Sherbrooke, Sherbrooke QC J1K 2R1, Canada
{arian.soltani,nkad2101,jordan.felicien.masakuna,marc.frappier, pierre-martin.tardif,froduald.kabanza}@usherbrooke.ca

Abstract. In light of the significant strides made by large language models (LLMs) in the field of natural language processing (NLP) [5], our research seeks to evaluate and contrast their proficiency in establishing associations within the realm of cybersecurity. Our experimental framework involves juxtaposing actual connections from various cybersecurity knowledge graphs (including MITRE CAPEC, D3FEND, and CVE connections to ATT&CK) against predictions made by LLMs using semantic textual similarity (STS). These connections span a broad spectrum, encapsulating diverse abstractions of threat descriptions, attack patterns, defense strategies, and vulnerabilities. The language models chosen for this study are varied, comprising state-of-the-art models from STS leaderboards, LLMs (GPT3.5 and PaLM), and ATTACK BERT [1], a cybersecurity domain-specific language model. Our experiments provide valuable insights into the differentiation between language models and data sources, thereby facilitating the broader application of STS in cybersecurity.

Keywords: Cybersecurity · Language Models · Intrusion Detection Systems (IDS) · MITRE ATT&CK

1 Introduction

The ever-increasing reliance on computer systems and digital technologies has led to a corresponding rise in cybersecurity threats, posing a significant risk to the security and integrity of cyber assets. As organizations increasingly depend on interconnected networks and digital infrastructures, the potential vulnerabilities have expanded, necessitating robust cybersecurity measures. Due to this trend, the demand for cybersecurity specialists has surged, reflecting the critical need for expertise in safeguarding sensitive information and defending against sophisticated cyber attacks. However, this growing demand is exacerbated by a pervasive talent shortage within the field of cybersecurity [6], presenting a challenge to organizations seeking to reinforce their cyber-defenses. Recent reports indicate that cybersecurity analysts overwork, leading to potential gaps in security coverage and increased vulnerability to cyber threats [17].

Artificial Intelligence (AI) has emerged as a crucial pillar in advancing cybersecurity, addressing the growing complexity of cyber threats. AI-powered tools can detect attacks in real-time, automating incident response processes and streamlining threat hunting. As the amount of data generated by connected devices increases exponentially, AI becomes essential in analyzing this data to mitigate cyber threats [16].

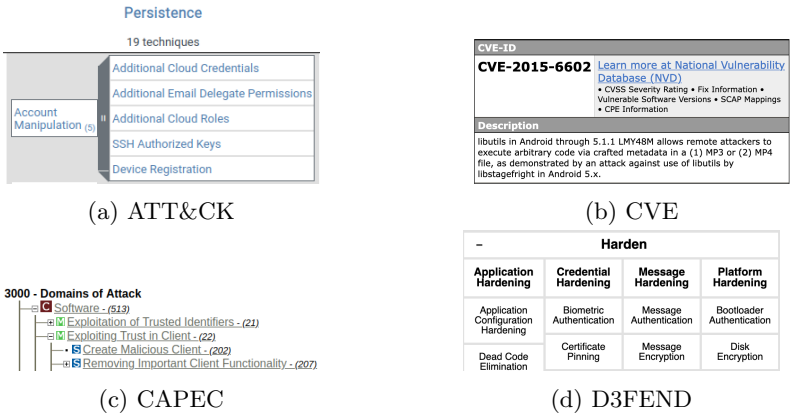
One of the most important ways AI can help alleviate the pressure on cybersecurity analysts is making associations among concepts and entities. Consider the following examples of tasks in the workflow of a cybersecurity analyst:

- finding the underlying goal of an attack (identifying the attacker's motives and objectives)
- predicting the next possible move of an attacker (threat prediction)
- suggesting defense mechanisms for threats (threat mitigation)
- anticipating possible threats based on vulnerabilities (proactive cyber risk management)

All of these tasks include recognizing **semantic relationships**, a capability that recent language models are increasingly adept at developing [13]. This notion has been previously explored for a narrow use-case, namely connecting vulnerabilities to threats [1], yet its general usability in the broader context remains uncharted. To our knowledge, currently there is no proposed general method of determining how competent AI systems are in making associations in the context of cybersecurity.

AI holds promise for enhancing cybersecurity efficiency, yet it faces challenges such as a lack of contextual understanding, leading to misclassifications. The necessity for transparency and interpretability in critical cybersecurity operations is not met by the opaqueness of large neural networks. Additionally, LLMs are prone to "hallucination", producing plausible but incorrect information [12], which undermines their reliability in cybersecurity contexts where accuracy is paramount. However, leveraging the embeddings from language models (LMs) could offer a compromise, harnessing LM strengths while mitigating generative AI weaknesses. Embeddings represent complex data as high-dimensional vectors, facilitating various algebraic operations. In cybersecurity, embeddings have been effectively applied to tasks like linking vulnerabilities to potential exploits, providing context-sensitive insights that are crucial for threat intelligence and incident response [1, 3]. Their compatibility with existing cybersecurity systems, such as intrusion detection systems (IDS) and security information and event management systems (SIEM), renders them a valuable tool for enhancing security measures.

Considering the advantages that embeddings offer for cybersecurity applications, our research focuses on evaluating the proficiency of current LMs in discerning a variety of semantic relationships pertinent to cybersecurity, including attack types, patterns, vulnerabilities, and countermeasures. We leverage relationships derived from cybersecurity knowledge bases to assess and compare the effectiveness of different LMs, thereby enabling the combination of various LMs to foster innovation in their application to cybersecurity tasks, such as ensemble



(a) ATT&CK (b) CVE

(c) CAPEC (d) D3FEND

Fig. 1: Examples from threat and defence description data sources

approaches. Preliminary findings underscore the promise of LM embeddings in identifying semantic connections within the cybersecurity domain.

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of the relevant literature and background information; Section 3 delineates the methodology employed in our study; Section 4 presents the results of our experimental evaluations; and the concluding section discusses implications for future research.

2 Background and Related Work

Background. Cybersecurity knowledge bases such as MITRE ATT&CK, CAPEC, D3FEND, and CVE play a pivotal role in the workflow of cybersecurity analysts, each covering an aspect of defense. *ATT&CK* enumerates and taxonomizes singular attack tactics, techniques, and procedures (TTPs), enabling analysts to perform cyber-threat intelligence (CTI), intrusion detection, risk assessment, and many other operations [15]. *CAPEC* catalogs known and common patterns of attack. In CAPEC, attack patterns are described and connected to related TTPs and weaknesses. *CVE* is the common vulnerability enumeration database, containing short descriptions of exposed vulnerabilities. Vulnerabilities are computer system weaknesses that enable attackers to take advantage. They are continuously discovered, exposed, exploited, and patched by developers, analysts, and attackers. *D3FEND* focuses on the opposite aspect compared to the aforementioned; it contains a top-down hierarchy of defensive strategies. Entries are linked to ATT&CK to mark the threats they can mediate. These knowledge bases contain cybersecurity domain knowledge and connections among them can be harbored as a testing ground for assessing LMs' capability in recognizing cybersecurity semantic similarities. First, previous cases of LM application in cy-

bersecurity are presented. Then in the next sections, the generalized comparison grounds are presented contrasting LMs' general capability in this domain.

Related Work. Language models present immense opportunities for applications in cybersecurity, owing to their ability in processing natural language content. This ability can be exploited for matching threats in emails, messaging apps, social media, event logs, vulnerability descriptions, etc. with high speed. Finding matches using LMs is achieved in three main ways: using a corresponding task in MLP, general purpose language models, and specialized language models. Wåreus and Hell introduced a method that can track vulnerabilities in software versions automatically [18] by utilizing named entity recognition (NER). Kuppa et al developed a joint word embedding space to match ATT&CK and CVE entries [11]. General purpose LMs for generating sentence embeddings such as Sentence BERT and its variations also proved useful in matching CAPEC and CVE entries [10], and expanding D3FEND-ATT&CK connections for incident response [3].

First endeavours for developing a domain-specific language models were fine-tuning BERT, a general-purpose base language model, on cybersecurity corpus using the masking technique [14, 2] optimized in performing NER and sentiment analysis, and capable of generating *word* embeddings. After, Abdeen et al introduced ATTACK BERT, a specialized cybersecurity *sentence* embedding model, and used it to match CVE and ATT&CK entries outperforming Sentence BERT. With the advent of LLMs, their generative capabilities are utilized [4, 7]. However, their embedding capabilities are not yet examined.

Introducing new concepts to cybersecurity where robustness and tangibilty are vital, requires investigation beyond narrow use-cases. By exploiting the information already used in operation, our approach introduces a testing ground encompassing several aspects of semantic connections in cybersecurity. This highlights the distinction of LMs for applications in cybersecurity and enables subsequent research to continuously improve, and tangibly infuse cybersecurity tools such as IDSs with the capabilities of modern LMs.

3 Methodology

3.1 Data

To assess LMs' capability in recognizing semantic connections in the context of cybersecurity, we extract descriptions and labelled connections among cybersecurity knowledge bases (ATT&CK, etc.), and then based on the description, LMs can predict relationships using STS. By comparing the ground truth with LMs' predictions, LMs' capability in recognizing these relationships can be measured.

The selected datasets are not exhaustive. There exist other semantic connections among entries, though the existing definitions and connections are rigorously reviewed by subject matter experts (SMEs) and the chosen sources are already being used by cybersecurity analysts. Therefore, we assume that the

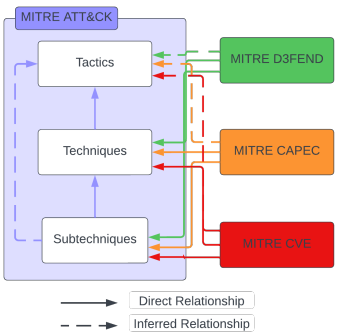


Fig. 2: Data source connections

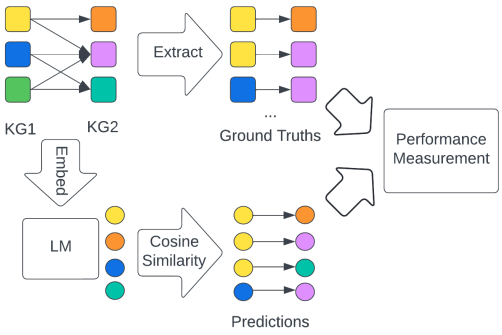


Fig. 3: Processing pipeline

outlined relationships are evident, and they must be unmistakable for a cybersecurity analyst.

After extracting, cleaning, and preprocessing definitions and relationships, definitions will be processed into embedding vectors using several language models. There are several ways to perform the preprocessing steps. For example, in [1], a subset of the descriptions matching the attack vector templates are selected. In contrast, our goal is to measure LMs ability in understanding and generalizing definitions. So instead of snippets, embeddings of the whole definition documents are processed in this work. Moreover, contrary to previous works, we expand the dataset to unambiguous adjacencies. For example, "CAPEC-25: Forced Deadlock" is mapped to subtechnique "T1499.004: Endpoint Denial of Service: Application or System Exploitation" in ATT&CK. In this instance, we expand the labels to the subtechniques parent nodes: technique "T1499: Endpoint Denial of Service", and tactic "TA0040: Impact". The validity of this expansion is supported by the hierarchical tree structure of ATT&CK. Fig. 2 depicts the connections among data sources, marking the extended (indirect) connections as dashed lines.

3.2 Similarity and performance measures

Having processed the documents to embeddings and extracted the ground truth, LMs' predictions can be measured. To achieve this using the common STS process, cosine similarity between the embedding vectors are calculated, and the top scoring vector pairs are selected (Fig. 3). For two vectors $\vec{a}$ and $\vec{b}$:

$$\text{Cosine Similarity}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

So assuming the embedding function $\vec{f_e}$ calculates embeddings of a text sequence, for two text sequences a and b, the similarity score would be:

$$\text{Similarity}(a, b) = \text{Cosine Similarity}(\vec{f_e}(a), \vec{f_e}(b))$$

Group	Subtask Name	Pair Count
Inter-ATT&CK	TecTac	227
	SubTec	411
	SubTac	558
CAPEC-ATT&CK	CapTac	271
	CapTec	235
	CapSub	156
D3FEND-ATT&CK	DefTac	795
	DefTec	1271
	DefSub	2631
CVE-ATT&CK	CveTac	2484
	CveTec	2057
	CveSub	498

Table 1: Pair counts for each subtask

In this study, similar to [1], we employ a similarity score to evaluate LMs' predictive capabilities through various multilabel classification evaluation metrics: recall@k, coverage error (CE), label ranking average precision (LRAP), and label ranking loss (LRL). Each metric serves as a distinct indicator of performance from various perspectives. Coverage error quantifies the average number of top predictions needed to encompass all actual labels, with its maximum value being the total number of labels. LRAP measures the percentage of top-ranked labels that are correct, indicating the precision of the ranking process. Label ranking loss (LRL) assesses the average ratio of incorrectly to correctly ranked labels, providing insight into the model's ranking accuracy. Lastly, recall@k determines the fraction of true labels that appear among the top-k predictions, offering a measure of the model's ability to capture relevant labels.

3.3 Language Models

Selected language models for this work are from the Massive Text Embedding Benchmark (MTEB) Leaderboard[8]. This list represents state of the art sentence embedding language models in several tasks such as classification, clustering, reranking, retrieval, STS along with information about the models' size, embedding dimensions, and maximum accepted tokens. It is continuously updated, introducing more advanced models outperforming previous state-of-the-art. Our curated selection features top-tier models from the MTEB leaderboard, alongside two distinguished Large Language Models (LLMs): GPT-3.5 (the foundational model behind ChatGPT) and PaLM. Additionally, we spotlight a specialized sentence embedding model tailored for the cybersecurity domain: ATTACK BERT (AB) [1]. Beyond these individual models, we also present a series of their equally weighted ensembles, which are denoted with an "E-" prefix.

4 Experimental Evaluation

To perform the comparison, first we prepare a dataset of pairs containing relations between knowledge base entries in section. Due to the variety of the data sources, and thus their different use-cases, the data is broken to several parts to highlight their difference, and they will be investigated as different subtasks. First, each of these subtasks will be explained, then the experiment results will be presented. Finally, results will be inspected in the corresponding section.

4.1 Subtasks

For our study, we collect data pertinent to specific subtasks from their respective sources, enabling us to evaluate language models' performance on individual subtasks (as shown in Fig. 4) and their combined average (refer to Table 2). Subtasks are designated based on cross-referencing elements from various cybersecurity frameworks: Tactics (Tac), Techniques (Tec), and Subtechniques (Sub) from ATT&CK; Patterns from CAPEC (Cap); Vulnerabilities from CVE (Cve); and Defense Strategies from D3FEND (Def). Within the MITRE ATT&CK framework, we examine the structural relationships—TecTac (Techniques to Tactics), SubTec (Subtechniques to Techniques), and SubTac (Subtechniques to Tactics)—to assess LMs' proficiency in matching abstract and technical concepts. We apply a similar approach to analyze the interplay between CAPEC attack patterns (CapTac, CapTec, CapSub), CVE vulnerabilities (CveTac, CveTec, CveSub), and D3FEND defense strategies (DefTac, DefTec, DefSub). These subtasks are critical for understanding various aspects of cybersecurity. This structured approach allows us to dissect and quantify how well LMs can discern and interpret the intricate web of relationships that define the cybersecurity landscape.

The number of pairs available for each subtask is displayed in Table 1. For these subtasks, measures (LRAP, coverage error, etc.) are calculated among predicted LM connections and ground truth labels, presented in Table 2. Moreover, in Fig. 4 for each subtask, a sweep on the k parameter of recall@k has been provided, demonstrates how LMs' precision changes at according to the number of top selected suggestions.

4.2 Results

Figure 4 aims to differentiate between subtasks, whereas Table 2 sheds light on the performance from the perspective of LMs. The analysis reveals significant variability in model effectiveness across subtasks and metrics, with ensemble models surpassing both domain-specialized models such as ATTACK BERT (AB) and LLMs.

Such fluctuations emphasize the impact that differing architectures, training methodologies, and datasets have on a model's ability to comprehend and process various concepts effectively. By incorporating this diversity, ensemble models enhance performance across various tasks. The recall@k metrics further reveal

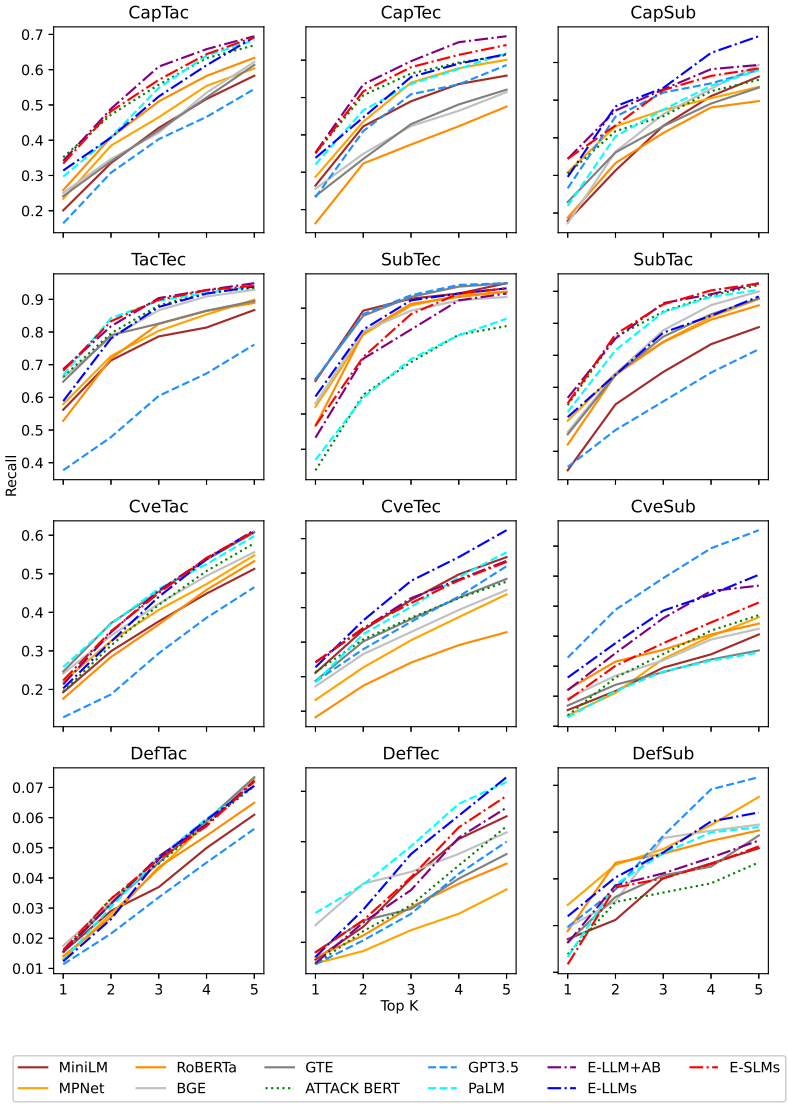


Fig. 4: Change of recall@k on top k parameter

Language Model	Variation	CE	LRAP%	LRL	Recall@5%
MiniLM	`all-MiniLM-L12-v2`	15.10	66.3%	0.123	43.9%
MPNet	`all-mpnet-base-v2`	15.18	68.0%	0.113	45.3%
RoBERTa	`all-roberta-large-v1`	15.85	65.8%	0.115	43.4%
BGE	`bge-large-en-v1.5`	15.37	68.3%	0.106	45.4%
GTE	`gte-large`	14.58	68.4%	0.103	45.2%
ATTACK BERT	`basel/ATTACK-BERT`	14.39	70.4%	0.096	46.7%
GPT3.5	`text-embedding-ada-002`	13.79	64.8%	0.132	42.5%
PaLM	`textembedding-gecko@001`	12.96	70.0%	0.090	47.3%
E-LLM+AB	GPT3.5 + PaLM + AB	12.31	**72.0%**	**0.085**	**48.7%**
E-LLMs	GPT3.5 + PaLM	**11.59**	70.7%	0.090	**48.7%**
E-SLMs+AB	BGE + GTE + AB	13.27	71.7%	0.088	48.4%

Table 2: Language models' average results

that there isn't a single model that consistently leads in performance across all subtasks, thereby underscoring the importance of selecting models tailored to specific tasks. For instance, GPT-3.5 shows prowess in discerning connections between subtechniques but falls short in associating related tactics. The performance in defensive tasks is particularly weak, with a maximum recall of 0.08 in the top 5 selections, revealing a substantial deficiency in current models' capacity to automate defensive cybersecurity tasks. This aligns with prior studies that have recognized the intricate and context-sensitive nature of these tasks [3, 9].

In other subtasks, the findings suggest that while the current LM embeddings might not reach complete autonomy in recognizing relationships, they may still play a crucial role in streamlining the decision-making process for analysts by helping prioritize options. Notably, smaller LMs are shown to deliver performance on par with their larger counterparts, offering a cost-effective solution for scenarios requiring the processing of large volumes of data where deploying LLMs may be financially impractical. Despite the intriguing insights these experiments offer, the occasional discrepancies in performance across different metrics signal a need for further investigation to confirm and refine these observations.

5 Conclusion and Future Work

This study evaluated LMs' effectiveness in identifying semantic relationships within the cybersecurity domain, leveraging established links between cybersecurity knowledge bases. Our findings indicate that singular and ensembled LMs possess a certain level of proficiency in executing specific tasks, potentially reducing the workload for cybersecurity analysts. However, when it comes to devising defensive strategies, further investigation is necessary. This research serves as a preliminary step, suggesting that incorporating additional data sources could enhance the robustness and reliability of the findings, thereby accelerating the integration of natural language processing (NLP) advancements into cybersecurity. Furthermore, the presented methodologies are poised for implementation in

current cybersecurity frameworks to provide practical value in real-world applications.

References

1. Abdeen, B., Al-Shaer, E., Singhal, A., Khan, L., Hamlen, K.: SMET: Semantic Mapping of CVE to ATT&CK and Its Application to Cybersecurity. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 243–260. Springer (2023)
2. Aghaei, E., Niu, X., Shadid, W., Al-Shaer, E.: SecureBERT: A Domain-Specific Language Model for Cybersecurity. In: International Conference on Security and Privacy in Communication Systems. pp. 39–56. Springer (2022)
3. Akbar, K.A., Halim, S.M., Hu, Y., Singhal, A., Khan, L., Thuraisingham, B.: Knowledge Mining in Cybersecurity: From Attack to Defense. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 110–122. Springer (2022)
4. Al-Hawawreh, M., Aljuhani, A., Jararweh, Y.: Chatgpt for cybersecurity: practical applications, challenges, and future directions. Cluster Computing **26**(6), 3421–3436 (2023)
5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023)
6. Crumpler, W., Lewis, J.A.: The Cybersecurity Workforce Gap. JSTOR (2019)
7. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access (2023)
8. Huggingface: MTEB Leaderboard. https://huggingface.co/spaces/mteb/leaderboard (2023), [Online; accessed 1-December-2023]
9. Kaiser, F.K., Andris, L.J., Tennig, T.F., Iser, J.M., Wiens, M., Schultmann, F.: Cyber threat intelligence enabled automated attack incident response. In: 2022 3rd International Conference on Next Generation Computing Applications (NextComp). pp. 1–6. IEEE (2022)
10. Kanakogi, K., Washizaki, H., Fukazawa, Y., Ogata, S., Okubo, T., Kato, T., Kanuka, H., Hazeyama, A., Yoshioka, N.: Tracing CVE Vulnerability Information to CAPEC Attack Patterns Using Natural Language Processing Techniques. Information **12**(8), 298 (2021)
11. Kuppa, A., Aouad, L., Le-Khac, N.A.: Linking CVE's to MITRE ATT&CK Techniques. In: Proceedings of the 16th International Conference on Availability, Reliability and Security. pp. 1–12 (2021)
12. McKenna, N., Li, T., Cheng, L., Hosseini, M.J., Johnson, M., Steedman, M.: Sources of Hallucination by Large Language Models on Inference Tasks. arXiv preprint arXiv:2305.14552 (2023)
13. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I., Roth, D.: Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Computing Surveys **56**(2), 1–40 (2023)
14. Ranade, P., Piplai, A., Joshi, A., Finin, T.: CyBERT: Contextualized Embeddings for the Cybersecurity Domain. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 3334–3342. IEEE (2021)

15. Roy, S., Panaousis, E., Noakes, C., Laszka, A., Panda, S., Loukas, G.: SoK: The MITRE ATT&CK Framework in Research and Practice. arXiv preprint arXiv:2304.07411 (2023)
16. Sarker, I.H., Furhad, M.H., Nowrozy, R.: AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. SN Computer Science **2**, 1–18 (2021)
17. Venturebeat: Mental Health: 66% of cybersecurity analysts experienced burnout this year. https://venturebeat.com/security/mental-health-cybersecurity-analysts/ (2023), [Online; accessed 19-July-2023]
18. Wåreus, E., Hell, M.: Automated CPE Labeling of CVE Summaries with Machine Learning. In: Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA 2020, Lisbon, Portugal, June 24–26, 2020, Proceedings 17. pp. 3–22. Springer (2020)