A Generative Artificial Intelligence Methodology for Automated Zero-Shot Data Tagging to Support Tactical Zero Trust Architecture Implementation

Charles Barzyk[1], Joseph Hickson[1], Jerik Ochoa[1], Jasmine Talley[1], Mikal Willeke[1], Sean Coffey[2], John Pavlik[2], and Nathaniel D. Bastian[1,2]

[1]Department of Systems Engineering, United States Military Academy, West Point, New York, 10996

[2]Army Cyber Institute, United States Military Academy, West Point, New York, 10996

Corresponding author's Email: nathaniel.bastian@westpoint.edu

Author Note: This work was supported by the U.S. Army Combat Capabilities Development Command C5ISR Center under Support Agreement No. USMA23011. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of Defense, or U.S. Government.

Abstract: A significant challenge in the implementation of the military's cybersecurity framework for Zero Trust Architecture (ZTA) is that the current approach for data tagging is done manually, which is a time-consuming and error-prone process that undermines the efficiency and effectiveness of cybersecurity measures. This paper introduces an innovative methodology that leverages generative artificial intelligence (AI) for automated data tagging to support tactical ZTA implementation within military mission command systems (MCS). Specifically, we develop the Generative AI ZEro-trust Labeling (GAZEL) tool, which uses a fine-tuned Large Language Model combined with in-context learning for automated zero-shot tagging of MCS message data according to predefined access control categories, streamlining the path towards an agile and fortified cybersecurity posture.

Keywords: Generative AI, Large Language Model, In-Context Learning, Data Tagging, Zero Trust Architecture, Cybersecurity

1. Introduction

In an era of increasingly rapid technology modernization, especially in the context of multi-domain military operations and digital warfighting, cybersecurity threats are becoming increasingly common (Cybersecurity and Infrastructure Security Agency, 2020; Keall, 2019; Woodiss-Field, Johnstone, & Haskell-Dowland, 2024). As a result, with more sophisticated attacks being created and used, the Department of Defense's (DoD) current cybersecurity architectures are becoming inadequate and data is increasingly being compromised from external sources. As such, the Federal Government has mandated the implementation of Zero Trust Architectures (ZTA) into governmental agencies in an attempt to improve cybersecurity posture, particularly within the DoD (*Executive Order 14028: Improving the Nation's Cybersecurity*, 2021; Young, 2022, 2021).

Traditional perimeter defense methods have proven inadequate as both external adversaries and insiders can bypass these defenses and access critical data without detection. ZTA fundamentally restructures this approach by eliminating the concept of a "trusted zone," applying stringent verification to all network connections and rigorously controlling access to data and resources throughout the organization (Army Cyber Center of Excellence, 2023). A major gap in ZTA implementation comes from the process of tagging data for use in ZTA, which falls into the 'data' pillar of ZTA frameworks. Data for use within ZTA is currently tagged manually, and so our proposed methodology herein aims to make this process more efficient by using generative artificial intelligence (GenAI) to automate the data tagging process to help enable tactical ZTA implementation.

Ongoing efforts seek to incorporate ZTA principles into the Army's current mission command systems (MCS), which aligns with the Army's Zero Trust vision and where most of the data breaches and cyber-attacks occur. Preceding research has narrowed our efforts on incorporating these principles into a tactical MCS, such as the Android Tactical Assault Kit for Military. We desire to optimize and bolster the Army's tactical data security to allow effective access control in combat environments. With the use of ZTA principles and the effective incorporation of GenAI to increase the efficiency of the data tagging process, it is possible to create a more lethal force through limiting our adversaries' and insiders' access to the military's information.

Given the existing technology gap for automated data tagging of tactical MCS data (Army Cyber Center of Excellence, 2023), specifically with Zero Trust (ZT) security in mind that can enable fine-grained access control and authentication decisions, the objective of this work is to proffer a novel methodological approach and develop an innovative analytical solution that utilizes GenAI for automated zero-shot tagging of tactical MCS data essential for ZTA implementation on the modern battlefield.

2. Literature Review

State of the art analytical tools, like LLMs, can enable cybersecurity practitioners to implement new technological defenses for access control, data security, and detection processes in order to respond to ongoing advances in cyber threats, particularly within the DoD (Kaur, Gabrijelčič, & Klobučar, 2023). Additionally, the evolution of network defense at the tactical level of a constrained, highly dynamic, and adversarial environment will necessitate the integration of ZT principles. Unlike traditional perimeter-based security models, ZT operates with continuous risk assessment and monitoring of network and user interactions. For ZT in tactical MCS, a crucial prerequisite for the implementation of ZT principles is the availability of tagged tactical data. This tagged data is the bedrock for enabling ZT applications such as user authentication, access control, and automated security responses. This presents a significant bottleneck, particularly in tactical environments where the volume and speed of data intake may require an automated approach to data tagging. Currently, data tagging for ZT applications requires extensive manual tagging, which is both costly and inefficient. Manual data tagging (classification marking) typically relies on application plug-ins to help humans apply banners, portion marks, and classification blocks consistently and expediently. Incorporating this process requires language to be consistent to ensure other humans can interpret markings consistently (ISOO, 2018). The National Security Agency is currently working to employ a new standardized ZT data format aimed at increasing the speed and generalization of automatic tagging (ISOO, 2018). However, they do not propose any ways to switch preexisting datasets into the new standardized format, requiring a manual tagging approach.

GenAI allows models to be discovered that represent joint probability distributions over the kinds of data encountered in artificial intelligence applications, such as natural images, audio wave forms containing speech, and symbols in natural language corpora. One GenAI technique is the Generative Adversarial Network, where a generative model is pitted against an adversary - a discriminative model - that learns to determine whether a sample is from the model distribution or the data distribution (Goodfellow et al., 2014). Foundation models represent a major shift in GenAI, emphasizing the development of large-scale, general-purpose models that serve as the base or "foundation" upon which more specialized functionalities and applications can be built (Bommasani et al., 2021). A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of down stream tasks (Bommasani et al., 2021). A Large Language Model (LLM) is a type of foundation model notable for its ability to achieve general-purpose language generation and other natural language processing tasks based on textual data. Zero-shot learning emerges as technique within machine learning where models are trained to handle tasks for which they have seen no examples during the training phase (Pourpanah et al., 2023). Fortunately, GenAI methods such as LLMs enable zero-shot learning and classification for data tagging without the need for manual tagging, offering a potential solution to the automation of data tagging. Recent experimental results have shown that LLMs have the potential to annotate data for different tasks at a relatively lower cost (Ding et al., 2023). Using LLMs combined with in-context learning enables the zero-shot, autonomous tagging of data without any training dataset. As such, this methodology may potentially be adapted to generate MCS data tags for tactical ZTA implementation.

The National Institute of Standards and Technology (NIST) explains the use of LLMs in application to cybersecurity as a structured approach to identify, protect, detect, respond to, and recover from cybersecurity threats and incidents (National Institute of Standards and Technology, 2018). NIST is looking at the application of LLMs to address prevalent problems in the cyber domain, such as intrusion detection. Despite this extensive application, a technology gap still exists for automating data tagging of tactical MCS data, specifically with ZT security in mind that can enable access control and authentication decisions in the DoD. In the context of LLMs for usage in cyber applications, Kucharavy et al. provide a brief overview of the history, state of the art, and implications of LLMs in terms of their principles and prospects in the context of Switzerland's cyber defense (Kucharavy et al., 2023). Their work is focused on user-based levels of cybersecurity such as having LLMs track Armed Forces members through artificial operator chat bots. This is similar to another work that states one advantage of LLMs is their ability to perform time-consuming and exhausting tasks without degradation of quality, or therefore automation (Barrett et al., 2023). These works are applying LLMs to areas we have not previously seen, but there has been no application to tactical ZT.

Existing research on utilizing LLMs for text annotation is unclear whether the automated data tagging workflows demonstrated in studies can be confidently applied to other datasets and tasks, given that they report minimal performance metrics and have only been tested on a small number of datasets (Pangakis, Wolken, & Fasching, 2023). Our work will be one of the first to analyze how large pre-trained foundation models (LPTMs) such as LLMs perform on automated annotation on tactical ZT data and not a public benchmark dataset. Our methodology uses a LLM fine-tuned with exercise data of tactical MCS message traffic between some devices to do zero-shot tagging of the data. This methodology will enable the creation of a capability that will be able to achieve and implement tactical ZTA for various datasets including those relating to the DoD.

3. Methodology

3.1 Stakeholder Analysis

As the first step of our stakeholder analysis, we identified 15+ Army and DoD organizations that had a vested interest in the implementation of ZTA, which included organizations such as the U.S. Army Cyber Command, U.S. Army Cyber Center of Excellence, U.S. Army Capabilities Manager - Cyber, U.S. Army Chief Information Officer, Program Executive Office - Intelligence, Electronic Warfare & Sensors, Program Executive Office - Enterprise Information Systems, and more. We then categorized the identified stakeholders by type (Decision Authority, Client, User, Owner, Interconnected), focusing on labeling the organization based on what we saw as their role in the ZTA implementation and technology modernization process, whether they are making the decisions on the requirements, implementing the product, or the organization requesting the product.

Once stakeholders were identified and categorized, we formed three focus groups to facilitate subsequent knowledge elicitation. Specifically, we identified a "Zero Trust Architectures" focus group to gain more information on the topic of ZT, particularly from a policy implementation perspective. Similarly, we identified a focus group on the topic of "Data Architectures and Strategies" in order to understand better how the Army and DoD process and stores data at scale at the enterprise and tactical levels. Finally, we identified a focus group on the topic of "Mission Command Systems" in order to identify the best tactical MCS data for tagging, as well as how ZT principles might fit into modernizing MCS data security. Once the focus groups were established, we created surveys tailored to each group to gather the information we needed from the stakeholders. Next, we executed virtual semi-structured interviews with the three focus groups in conjunction with providing them the surveys to complete. Upon consolidating the focus group insights, we produced a Findings, Conclusions, and Recommendations (FCR) matrix to begin the solution framing process to design our solution architecture diagram depicted in Figure 1.

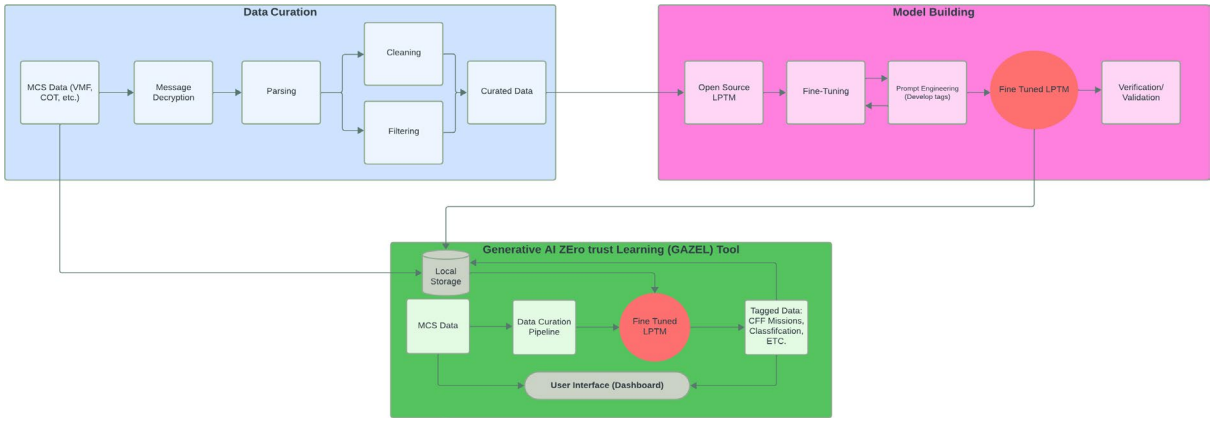

Figure 1: Solution Architecture Diagram

3.2 Solution Design and Development

LLMs are foundation models that are designed to understand, generate, and manipulate natural language. These models are trained on a large corpus of text data, enabling them to perform a wide range of language-related tasks. The fine-tuning process is a crucial stage in the development of LLMs. After an LLM is initially trained on a large, general dataset, it can be fine-tuned on a smaller, domain-specific dataset. This process allows the model to adapt its knowledge to specific industries, topics, or tasks, enhancing its performance and making it more relevant for targeted applications. Through fine-tuning and in-context learning, LLMs can be customized to meet the nuances of different requirements. In our case, by fine-tuning a LPTM on tactical MCS data, we can adapt the large knowledge base of the LLM to a smaller nuanced text corpus that it was not trained on. Through prompting, one can also provide an input into what the model should adapt to through natural language.

3.2.1 Data Curation

The data curation pipeline of our solution architecture, as depicted in Figure 1, begins with uploading the MCS data. After ingest, the data is first decoded and then it is parsed, cleaned and filtered in parallel to produce curated data to be used in the model building pipeline. First, we used two XML datasets from different Army exercises. The EDGE23 data was Cursor-on-Target (CoT) data for a training center exercise at Fort Irwin, California, whereas the second data set was from the Joint

Automated Deep Operations Coordination System (JADOCS) to Advanced Field Artillery Tactical Data System (AFATDS), providing Variable Message Format (VMF) data for communication between the two MCS. This data was parsed into messages as shown in Figure 2, and 600 messages were then manually tagged into one of three message categories into a single dataset. Messages were tagged under three types of access control for ZT, which were not present as XML tags and would not be able to be directly parsed from XML. These include: "Direct Message," which are messages that include information critical to a specific user such as unit location or passwords; "General Message," for content that is non-specific but must be logged if viewed by outsiders such as ammunition type requested or mission summaries; and "Outsider Message," for content unrelated to the mission that could be managed by outsiders without logging such as admin messages.

This dataset was then partitioned using stratified sampling into a validation set of 20% of the data in order to validate model results. Since the model was over-sampling the general message category, we resampled from the minority classes to balance out the training set. During data curation, we employ tokenization, a process where text is divided into tokens, which are fundamental linguistic units or strings of characters present within a text. For instance, the word "Cowboy" can be split into the tokens "cow" and "boy." This method allows the LLMs to grasp the statistical relationships between these tokens, thereby enhancing its ability to predict the subsequent token in a sequence (Wolfram, 2023). In our pipeline, tokens are assigned an end-of-sentence token as its padding token, meaning that our tokenizer will use an end-of-sentence token to fill in any extra spaces when aligning sequences to the same length. This is necessary to feed unique data types into the fine-tuning process. The messages are ran through this tokenizer as a last step before implementing them into the fine-tuning process.

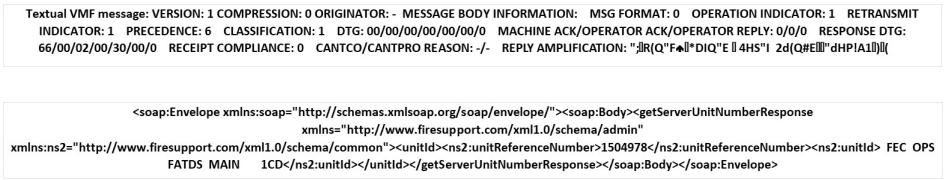

Figure 2: MCS Message Examples; VMF (Top), CoT (Bottom)

3.2.2 Model Building

In the model building pipeline, as can be seen in Figure 1, we focused on refining the capabilities of a LPTM by preparing tactical MCS data for processing, utilizing predefined tags identified from the XML headers and types of messages exchanged. This involved employing two open-source LLMs from Hugging Face: Bidirectional Encoder Representations from Transformers (BERT) and TinyLLAMA (Devlin, Chang, Lee, & Toutanova, 2019; Zhang, Zeng, Wang, & Lu, 2024). Both models are built off of the self-attention mechanism of the Transformer (Vaswani et al., 2017), which have a self-attention mechanism that allows models to weigh the importance of different parts of an input sequence when producing an output. For fine-tuning these LLMs, we opted for a Parameter-Efficient Fine Tuning (PEFT) strategy, specifically leveraging the Low-Rank Adaptation (LoRA) method. LoRA is designed to minimize computational requirements by maintaining the original model weights unchanged and introducing trainable rank decomposition matrices within each transformer layer (Hu et al., 2021). This technique significantly lowers the parameter count needed for fine-tuning, allowing for more efficient model adaptation with a only a set of 600 manually-tagged messages. To enhance model accuracy, we used prompt engineering; we devised the following prompt through experimentation to guide the LLM in classifying (tag) messages into one of three ZT access control categories:

"Classify the message using these tags: Direct Message = message intended for a specific user, information critical for a specific user only. General Message = message is generic could be viewed/reported on by anyone in system, must be logged if viewed by an outside. Outsider Message = message is unrelated to mission and can be successfully reported by an outsider with no log."

Following the application of PEFT using the LoRA method and prompt engineering, the LPTMs were customized on the fine-tuned MCS data. BERT required only 10 seconds for this training phase, while TinyLLAMA took substantially longer, at 2.5 hours. Subsequently, both models underwent a validation phase using a set of manually tagged messages to test their tagging accuracy and classification capabilities into one of the three ZT access control categories.

3.2.3 Generative AI ZEro-Trust Labeling (GAZEL) Tool

We developed GAZEL, with the User Interface (UI) depicted in Figure 3, that allows the user to quickly upload de-coded, cleaned MCS data (CoT and VMF) that is then subsequently auto-tagged using our zero-shot GenAI-based methodology. GAZEL includes the option to choose the model (BERT or TinyLLaMa) while displaying simple statistics (a histogram showing the occurrence of each of the three tag categories) and the ability to download a CSV file of the tagged data. The tool's UI was built using the Plotly Dash package for Python as it was easy to containerize for subsequent deployment by the end-user.

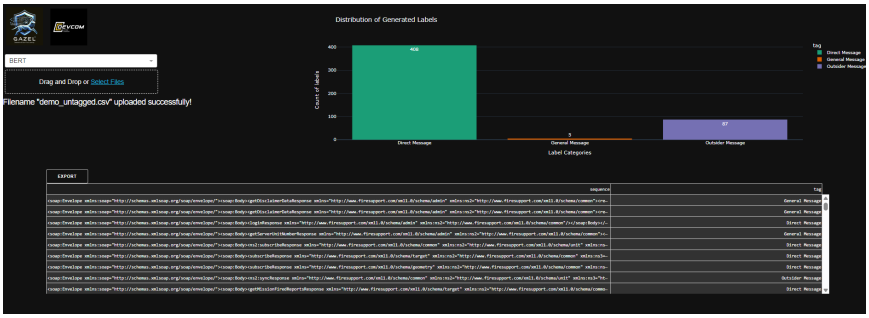

Figure 3: GAZEL User Interface

4. Computational Experimentation

We conducted rigorous computational experimentation for model evaluation. Table 1 presents the outcomes for two distinct LLMs. The comparison indicates that BERT achieved superior performance in automated data tagging, demonstrating both higher accuracy and more efficient fine-tuning times. With 110M parameters, BERT is significantly leaner compared to TinyLLama, which boasts 1.1B parameters. This substantial difference in model size accounts for the longer fine-tuning duration required for TinyLLama. Despite its larger parameter count, TinyLLama did not surpass BERT in accuracy using the same fine-tuning and prompt engineering techniques. BERT notably achieved an accuracy of 0.78, underscoring its effectiveness. While we only experimented with BERT and TinyLLama models shown in Table 1, our methodology is adaptable to any LPTM.

Table 1: Comparison of Validation Results for BERT and TinyLLaMA Models

Metric	BERT	TinyLLaMA
Evaluation Loss	**0.842**	0.927
Evaluation Accuracy	**0.781**	0.593
Evaluation Runtime (seconds)	**0.282**	30.4
Fine-Tuning Runtime (seconds)	**11**	9,420

5. Conclusion

To effectively implement a ZTA security framework in tactical environments, the automation of data tagging is essential. The conventional manual approach to data tagging, characterized by its labor-intensive, error-prone nature, falls short in meeting the efficiency and accuracy demands of modern security. The advent of GenAI, particularly through advancements in LLMs, introduces a pivotal capability for analyzing extensive text inputs and generating coherent responses grounded in vast textual knowledge bases. We demonstrate in our research the potential of leveraging a fine-tuning methodology, which involves a compact subset of data coupled with strategic prompt engineering, to significantly enhance data tagging efficiency.

In the context of tactical MCS data, our application of a finely-tuned BERT model, augmented with LORA PEFT and straightforward prompt engineering techniques, demonstrates a notable 78% accuracy rate in data tagging. This underscores the feasibility of our methodology, despite the constraints of subjectively defined data tagging categories for these messages. Such results not only showcase the model's capability to generate highly accurate tags under nuanced categorization but also highlight the potential of our tool, GAZEL. GAZEL represents a first step towards fulfilling the gap for automated data tagging solutions essential for the enforcement of ZT principles.

However, the potential for our proffered GenAI methodology and innovative analytical tool does not end at this first iteration. Our research herein encapsulates a reproducible methodology able to be applied to future data tagging efforts for DoD ZTA implementation. Future work will explore the use of other base LPTMs, as well as other fine-tuning and in-context learning approaches, such as retrieval augmented generation. Future enhancements to GAZEL will aim to give users the ability to directly input prompts and a select set of manually tagged data for real-time fine-tuning of the LLM via the UI. This advancement will enable users to tailor their data tagging efforts to specific needs within tactical environments, facilitating a more comprehensive approach to automated ZT data tagging. We will also work to develop an application programming interface (API) to enable real-time, streaming MCS data tagging without use of GAZEL's UI.

7. References

Army Cyber Center of Excellence. (2023). *Zero trust tactical implementation guide* (Technical Report). Fort Gordon, Georgia: Cyber Center of Excellence, Fort Gordon. (Controlled by: Army CCoE. CUI Category: CTI, OPSEC. Dissemination Control: FEDCON.)

Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., … Yang, D. (2023). Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, *6*(1), 1–52. doi: 10.1561/3300000041

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. *ArXiv*, *abs/2108.07258*. Retrieved from https://api.semanticscholar.org/CorpusID: 237091588

Cybersecurity and Infrastructure Security Agency. (2020). Ed 21-01: Mitigate solarwinds orion code compromise. Retrieved from https://www.cisa.gov/news-events/directives/ed-21-01-mitigate-solarwinds-orion-code-compromise.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., & Bing, L. (2023). *Is gpt-3 a good data annotator?*

Executive order 14028: Improving the nation's cybersecurity. (2021). Retrieved from https://www.gsa.gov/technology/ it-contract-vehicles-and-purchasing-programs/technology-products-services/it-security/executive-order-14028

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (p. 2672–2680). Cambridge, MA, USA: MIT Press.

Hu, J. E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *ArXiv*, *abs/2106.09685*. Retrieved from https://api.semanticscholar.org/CorpusID:235458009

ISOO. (2018). *Marking classified national security information.* Retrieved from https://www.archives.gov/files/ isoo/training/marking-booklet-revision.pdf

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. doi: https://doi.org/10.1016/j.inffus.2023.101804

Keall, C. (2019). *Experts warning as microsoft restores office 365 after worldwide outage.* Retrieved from https://www.nzherald.co.nz/business/experts-warning-as-microsoft-restores-office-365-after -worldwide-outage/JZVIHQJZSLT3WPXTEAXBVROSGQ/?c_id=3&objectid=12286870

Kucharavy, A., Schillaci, Z. M., Mar'echal, L., Wursch, M., Dolamic, L., Sabonnadiere, R., … Lenders, V. (2023). Fundamentals of generative large language models and perspectives in cyber-defense. *ArXiv*, *abs/2303.12132*. Retrieved from https://api.semanticscholar.org/CorpusID:257663521

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity. Retrieved from https://nvlpubs.nist.gov/nistpubs/cswp/nist.cswp.04162018.pdf

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative ai requires validation. *ArXiv*, *abs/2306.00176*. Retrieved from https://api.semanticscholar.org/CorpusID:259000016

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., … Wu, Q. M. J. (2023). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4051-4070.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wolfram, S. (2023, Feb). *What is chatgpt doing ... and why does it work?* Retrieved from https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (Accessed: 2023-03-08)

Woodiss-Field, A., Johnstone, M. N., & Haskell-Dowland, P. (2024). Examination of traditional botnet detection on iot-based bots. *Sensors*, *24*(3). Retrieved from https://www.mdpi.com/1424-8220/24/3/1027 doi: 10.3390/s24031027

Young, S. D. (2021). Improving detection of cybersecurity vulnerabilities and incidents on federal government systems through endpoint detection and response. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2021/10/ M-22-01.pdf

Young, S. D. (2022). Moving the u.s. government toward zero trust cybersecurity principles. Retrieved from https:// www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). *TinyLlama: An open-source small language model.*