NLP-based automated scoring of OT misconfigurations via CWE and CVSS mapping

NLP-based automated scoring of OT misconfigurations via CWE and CVSS mapping

1st Mario Todaro
Politecnico di Torino
Torino, Italy
mario.todaro@polito.it

2nd Alberto Salvatore Colletto
AlphaWaves S.r.l.
Torino, Italy
a.colletto@awaves.it

3rd Alessio Viticchié
AlphaWaves S.r.l.
Torino, Italy
a.viticchie@awaves.it

4th Alessandro Aliberti
Politecnico di Torino
Torino, Italy
alessandro.aliberti@polito.it

***Abstract*—Misconfigurations within Operational Technology (OT) environments represent a significant source of cyber risk, often resulting in critical disruptions to industrial processes. However, the absence of standardized methodologies for quantifying their impact hinders effective risk assessment and prioritization. This study proposes a novel and fully automated framework that maps misconfigurations to the Common Weakness Enumeration (CWE) taxonomy through semantic similarity techniques, employing state-of-the-art sentence embedding models and cosine similarity metrics. The framework enables the computation of quantitative risk indicators by linking the identified CWEs to associated Common Vulnerabilities and Exposures (CVEs) and aggregating their Common Vulnerability Scoring System (CVSS) scores. A voting ensemble of pre-trained language models is introduced to enhance robustness and semantic accuracy. Experimental validation demonstrates improved precision over single-model baselines, confirming the efficacy of the proposed approach. The resulting system offers a scalable, data-driven tool for OT stakeholders to evaluate and prioritize misconfiguration-related cybersecurity threats systematically.**

***Index Terms*—Cybersecurity, Misconfiguration Risk, Natural Language Processing, Sentence Embeddings, Operational Technology.**

I. Introduction

The proliferation of innovative and highly interconnected IT solutions, enabled by advances in cloud computing, virtualization, and ubiquitous connectivity, has expanded operational capabilities across sectors. However, this digital transformation has also significantly increased the attack surface, creating new vulnerabilities that malicious actors can exploit. In enterprise and industrial contexts, particularly those involving Operational Technology (OT), the implications of cyber-attacks are especially critical [1]. A single breach can lead to substantial economic losses, reputational damage, or the compromise of sensitive intellectual property and strategic infrastructure [2]. As a result, cybersecurity has become a central pillar of organizational resilience and continuity.

Among the diverse set of threat vectors, system misconfigurations represent a particularly pervasive and insidious risk often resulting from human error, such as misconfigured access controls, services, or software parameters. While they may not constitute vulnerabilities in the strict technical sense, they often serve as enablers for exploitation. Despite their importance, misconfigurations are difficult to evaluate systematically due to the absence of standardized risk assessment frameworks that can translate high-level security guidelines into quantifiable indicators.

To address this gap, this work introduces a novel, fully automated architecture for the semantic evaluation and scoring of misconfigurations in OT environments. The proposed methodology leverages recent advances in natural language processing, specifically sentence embeddings and cosine similarity, to map unstructured best practice descriptions to formalized entries within the Common Weakness Enumeration (CWE) taxonomy. This semantic alignment enables the identification of conceptually similar weaknesses, thus facilitating structured reasoning about risks that were previously defined only in informal terms. The architecture further integrates these mappings with vulnerability data by linking each identified CWE to associated entries in the Common Vulnerabilities and Exposures (CVE) database and aggregating their corresponding Common Vulnerability Scoring System (CVSS) scores. The resulting framework provides a quantitative, interpretable risk score that supports more objective prioritization of mitigation efforts. To improve robustness and reduce sensitivity to linguistic variability, the system employs an ensemble of state-of-the-art sentence embedding models, enhancing the accuracy and reliability of semantic associations. Experimental validation demonstrates that the proposed ensemble-based approach outperforms single-model baselines in assigning CWEs that are thematically aligned with relevant best practices. By transforming abstract security guidelines into actionable and measurable insights, this work provides a practical decision-support tool for risk assessment in OT context. In doing so, it advances the state of the art in automated cybersecurity evaluation and contributes a reproducible methodology for bridging the gap between operational recommendations and formal risk metrics.

The rest of the paper is organized as follows: Section II provides an overview of the background and state of the art, focusing on cybersecurity scoring frameworks, misconfiguration assessment, and semantic similarity techniques in natural language processing. Section III introduces the proposed architecture, describing the methodology for semantic alignment

This publication is partially part of the project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3-SERICS PE 00000014 – CUP F53C22000740007.

between best practices and CWE entries, the ensemble modeling strategy, and the computation of risk scores through CVE-CVSS aggregation. Section IV presents the experimental setup and evaluation criteria, analyzing the framework performance across multiple cybersecurity categories. Finally, Section V discusses the results, highlighting key findings, limitations, and opportunities for methodological refinement.

II. Background and State of the Art

This section provides the conceptual and technical background necessary to support the design of the proposed methodology. It covers key topics from the domains of cybersecurity and AI, including vulnerability classification standards, risk scoring frameworks, and natural language processing techniques, all of which underpin the automated risk assessment strategy presented in this work.

A. Taxonomies of Security Flaws and Misconfigurations

A fundamental step in assessing and mitigating cyber threats involves understanding the classification of software and hardware flaws. Three distinct yet interrelated categories: weaknesses, vulnerabilities, and misconfiguration, play a central role in the security posture of modern digital systems. The CWE is a standardized taxonomy maintained by the MITRE Corporation that categorizes general classes of software and hardware weaknesses which, if unaddressed, may be exploited by attackers [3]. Each CWE entry provides a descriptive definition of a recurring flaw type (e.g., buffer overflows, improper access control), enabling developers, analysts, and security practitioners to design countermeasures early in the software development lifecycle. By structuring weaknesses hierarchically and thematically, CWE facilitates systematic analysis and supports proactive, design-time security interventions.

Complementing the CWE is the CVE system, which focuses not on general flaw types, but on specific, documented instances of vulnerabilities observed in operational environments [4]. Also maintained by MITRE in collaboration with the U.S. Department of Homeland Security, the CVE database assigns unique identifiers to vulnerabilities, accompanied by textual descriptions, discovery timelines, and references to exploited CWE classes. Each CVE provides actionable threat intelligence by pinpointing real-world weaknesses that have been disclosed or exploited, often linking directly to vendor advisories and patches.

In contrast to CWEs and CVEs, misconfigurations are not formalized within a centralized taxonomy. Misconfigurations arise when system components, whether hardware, software, or network elements are configured in ways that deviate from secure practices, typically due to human error, inadequate documentation, or oversight [5]. Although misconfigurations do not constitute inherent weaknesses within the codebase, they can still expose critical assets to exploitation. Addressing misconfigurations remains particularly challenging due to the absence of standard identifiers and the lack of representation in publicly maintained repositories. Their mitigation relies instead on compliance with evolving best practices and adherence to guidelines published by cybersecurity authorities and relevant industrial bodies.

B. Cybersecurity Scoring Systems

In the field of cybersecurity, risk quantification is supported by several standardized metrics developed under the guidance of the National Institute of Standards and Technology (NIST) [6]. Among these, the *Common Vulnerability Scoring System (CVSS)* is the most widely adopted framework for evaluating the severity of known vulnerabilities. CVSS assigns each vulnerability a score ranging from 0 to 10, based on two principal components: impact (which reflects the potential compromise to confidentiality, integrity, and availability) and exploitability, which measures the ease with which an attack can be carried out.

The *Common Configuration Scoring System (CCSS)* extends this concept to assess the risk associated with misconfigurations in hardware and software systems. While the underlying model has been proposed, the CCSS framework remains under active development and requires further refinement before it can be widely implemented.

Similarly, the *Common Misuse Scoring System (CMSS)* is designed to evaluate the risks arising from incorrect or unintended use of software and hardware. Like CCSS, CMSS is in a preliminary stage and has not yet reached full maturity for operational deployment in enterprise security assessments.

C. Sentence Embeddings and Similarity Computation

To enable a rigorous and systematic computation of risk scores for conceptually abstract misconfigurations, the proposed architecture integrates deep learning techniques, specifically through the use of sentence embedding models. Sentence embedding is a Natural Language Processing (NLP) technique that transforms textual data into continuous, high-dimensional vector representations, thereby enabling the application of mathematical operations to capture semantic relationships. These dense vector representations are employed in a wide range of tasks, including semantic understanding, machine translation, question answering, and document clustering.

Among the most effective models for generating sentence embeddings are those from the Sentence-Transformers framework, a derivative of the BERT architecture [7]. These models utilize pre-trained transformers to tokenize input sentences while preserving both syntactic structure and semantic nuance. A key feature of Sentence-Transformers is their implementation of a Siamese network architecture, which significantly enhances their efficiency and accuracy in semantic similarity tasks. This makes them particularly well-suited for applications requiring the comparison of large volumes of text, such as the semantic mapping of misconfigurations to security-related knowledge bases.

Once textual inputs are converted into embedding vectors, semantic similarity can be evaluated using various distance metrics. Among these, cosine similarity is especially appropriate due to its sensitivity to orientation rather than magnitude. It

measures the cosine of the angle between two vectors, focusing on semantic content while remaining invariant to sentence length. The metric is defined as follows:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{|A||B|} \tag{1}$$

The resulting value ranges from 1 to -1, where 1 indicates identical semantic content and -1 represents complete dissimilarity.

D. Evaluation of Semantic Similarity Using the STS Benchmark

The *Semantic Textual Similarity Benchmark (STS Benchmark)* is a widely adopted evaluation framework designed to assess the ability of Natural Language Processing (NLP) models to quantify semantic similarity between sentence pairs [8]. Developed as part of the SemEval shared tasks, the benchmark provides a standardized, reproducible methodology for comparing the effectiveness of different semantic similarity techniques.

a) Dataset and Scoring: The benchmark includes a collection of sentence pairs, each annotated with a similarity score on a scale from 0 (completely dissimilar) to 5 (semantically equivalent), based on human judgment. The dataset is divided into *training*, *development*, and *test* sets, enabling robust model training and generalization across various textual domains and syntactic structures [8].

b) Evaluation Metrics: Performance is typically evaluated using the *Pearson* and *Spearman rank correlation coefficients*, which assess the degree of alignment between the similarity scores predicted by the models and those assigned by human annotators. Higher correlation values indicate better semantic fidelity and model robustness in capturing nuanced meaning relationships [8].

E. Proposed Architecture for Automated Misconfiguration Risk Assessment

Despite the availability of standardized taxonomies such as CWE and CVE, and scoring systems like CVSS, existing approaches to cybersecurity risk quantification still fall short when addressing the complexity of misconfigurations, particularly within OT environments. Current frameworks, including the CCSS and the CMSS, remain incomplete and lack wide-scale operational deployment. Moreover, most prior work in risk assessment has focused on known vulnerabilities rather than on configuration weaknesses and typically requires expert manual intervention or rule-based matching, which limits scalability and adaptability.

In contrast, the approach proposed in this work introduces a fully automated architecture that leverages recent advancements in NLP, specifically sentence embeddings and semantic similarity metrics—to bridge the gap between informal best practices and formalized cybersecurity taxonomies. Unlike traditional methods that rely on direct string matching or expert-encoded rules, the system performs semantic mapping from unstructured text (i.e., best practices) to CWE entries, which are then linked to associated CVEs and scored via CVSS. This enables the quantification of abstract and high-level security guidelines without requiring manual annotation or domain-specific heuristics.

A key advancement over the state of the art is the use of an ensemble of sentence embedding models, which improves robustness and reduces sensitivity to linguistic variance. This ensemble-based design, validated against domain-specific best practices, outperforms single-model baselines and demonstrates that combining diverse NLP models significantly enhances semantic alignment in a security context. Furthermore, the proposed architecture introduces a weighted risk computation mechanism that integrates similarity scores into CVSS aggregation, thereby producing a context-aware, interpretable, and scalable risk score. This is a notable step forward in enabling automated risk assessment for misconfiguration, an area often overlooked in both academic literature and practical deployment scenarios. By synthesizing formal security ontologies with modern NLP techniques and a lightweight computational framework, this work delivers a replicable and extensible foundation that addresses the core limitations of existing approaches, advancing the state of the art in OT cybersecurity risk evaluation.

III. Methodology

The primary objective of the proposed architecture is to establish a systematic connection between abstract concepts, such as best practices, typically formulated as general security guidelines and formally defined entities, such as CVEs and their associated Common CVSS metrics. Due to the semantic and structural disparity between these domains, the CWE framework serves as an effective intermediary. While CVEs provide highly specific and technical descriptions, often including vendor- or version-specific details, CWE entries encapsulate generalized classes of weaknesses that align more closely with the conceptual scope of best practices. The architecture is therefore grounded in a semantic similarity analysis between textual descriptions of best practices and the CWE corpus. This intermediate mapping enables the identification of related vulnerabilities and facilitates the computation of an aggregated CVSS-based risk score, thus transforming qualitative guidelines into quantitative and actionable security metrics.

A. Model Selection

To compute semantic similarity between best practices and CWE descriptions, several pre-trained sentence embedding models were considered. The final selection included five models from the Sentence-Transformers framework:

- `all-mpnet-base-v2`
- `all-MiniLM-L6-v2`
- `all-MiniLM-L12-v2`
- `paraphrase-multilingual-MiniLM-L12-v2`
- `paraphrase-MiniLM-L6-v2`

The selection process was guided by the need to balance three primary criteria: average task performance (measured across

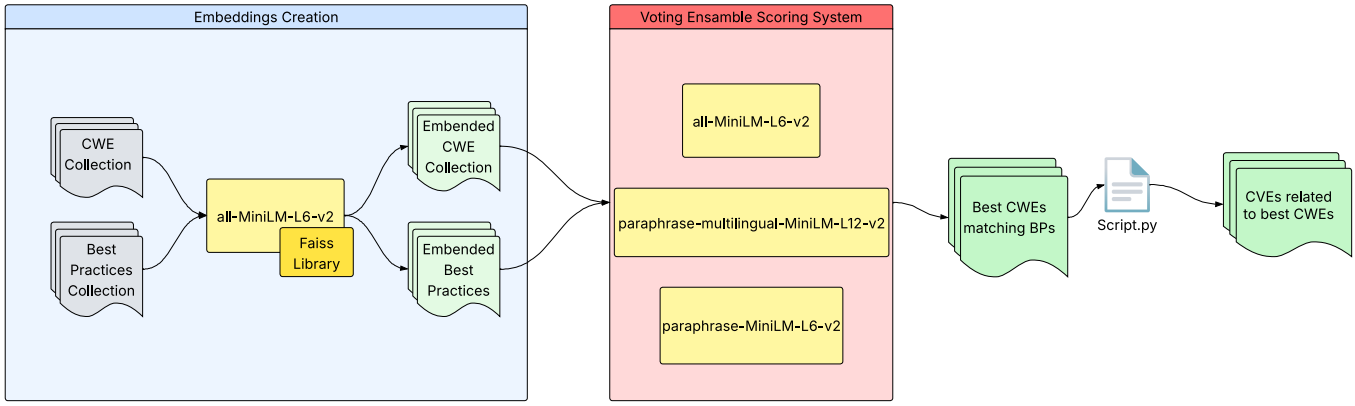

Fig. 1. Overview of the proposed risk scoring architecture. The system uses sentence embeddings and a voting ensemble to map best practices to CWEs, and queries-related CVEs, and aggregates CVSS scores to compute a final misconfiguration risk score.

standard semantic search and sentence embedding benchmarks), inference speed (measured in sentences per second on a V100 GPU), and model size (measured in megabytes). These criteria were chosen to ensure a trade-off between semantic accuracy, computational efficiency, and resource constraints.

Given the absence of a dedicated dataset for mapping best practices to CWE entries, and the lack of a definitive metric for identifying the most suitable model for this specific task, an initial evaluation was conducted using the Semantic Textual Similarity (STS) Benchmark [8]. This benchmark serves as a de facto standard for assessing semantic similarity in NLP and provides a reliable reference for model comparison. All five models demonstrated strong performance, achieving Pearson and Spearman correlation coefficients between 0.82 and 0.84, thus confirming their reliability for sentence-level semantic comparison.

To further differentiate among models, an additional empirical analysis was performed by computing the range of cosine similarity scores each model produced across a representative set of comparisons. The underlying hypothesis is that a more effective model should exhibit a greater distinction between high- and low-relevance CWE candidates when matched to a given best practice. A wider score distribution suggests a stronger ability to discriminate between semantically related and unrelated entities.

The evaluation was performed on a curated set drawn from the authoritative guide *Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies* [9]. This document serves as a widely recognized reference in the domain of industrial control system (ICS) cybersecurity and was selected to ensure domain relevance and practical applicability.

B. Voting Ensemble Scoring System

To enhance the reliability of risk score computation, the proposed architecture incorporates a voting ensemble of multiple sentence embedding models. While the use of multiple models does not inherently guarantee perfect accuracy, it significantly improves robustness from a probabilistic standpoint. This strategy is grounded in the principle of model diversity: by leveraging models with differing architectures and training datasets, the ensemble mitigates individual model biases and compensates for potential weaknesses in specific semantic representations. The decision to include three models reflects a deliberate balance between computational efficiency and evaluation quality. Preliminary analysis indicated that combining only two models yielded marginal improvements over single-model performance, whereas integrating more than three models introduced substantial computational overhead, adversely affecting scalability and usability in practical settings. To operationalize this approach, a structured workflow was developed that integrates best practice embeddings, CWE matching, and CVE-CVSS aggregation through an ensemble of embedding models. An overview of this process is illustrated in Figure 1, which outlines the main components involved in risk scoring, from semantic similarity computation to final risk estimation.

The final ensemble comprises the three models that exhibited the greatest variability in cosine similarity scores an indicator of discriminatory capacity when matching best practices to CWE descriptions:

- `paraphrase-multilingual-MiniLM-L12-v2`
- `all-MiniLM-L6-v2`
- `paraphrase-MiniLM-L6-v2`

This ensemble configuration enables a more nuanced and reliable estimation of semantic similarity, which is critical for accurately identifying relevant CWEs and computing associated risk scores.

C. Risk Score Computation

Each model within the ensemble independently computes semantic similarity scores by comparing each best practice to all descriptions in the CWE corpus. The resulting scores from each model are normalized by assigning equal weights—specifically, one-third per model ensuring a balanced contribution to the overall assessment. For each best practice, CWE entries are ranked based on their aggregated similarity scores, and the top five matches are selected.

The identifiers of these top-ranking CWEs are then used to query a structured database that maps each CWE to its associated CVEs. Subsequently, the CVSS values corresponding to each linked CVE are retrieved. To reflect the semantic relevance of each CWE, the similarity scores are employed as weighting factors in computing a weighted average of the CVSS values. This final weighted average constitutes the risk score, providing a quantifiable and context-aware assessment of each misconfiguration.

IV. Experimental Results

Before evaluating the proposed methodology, it is necessary to establish a framework for interpreting the results. To this end, a set of seven cybersecurity categories has been defined to facilitate the consistency assessment between the CWEs identified by the system and the 25 best practices used for testing. These categories reflect key functional domains within cybersecurity and serve as a reference taxonomy for classification. Given the general nature of both CWE descriptions and best practices, individual entries may justifiably belong to more than one category. The defined categories are as follows:

1) *Authentication & Access Control*: Encompasses mechanisms for verifying user identities and controlling access to system resources.
2) *Data Encryption*: Focuses on safeguarding sensitive information through appropriate cryptographic techniques.
3) *Security Monitoring & Incident Response*: Includes capabilities for threat detection, monitoring, and management of security incidents.
4) *Secure Software Design & Network Architecture*: Addresses the integration of security principles from the design phase of software and network systems.
5) *Patch Management*: Pertains to the identification, testing, and deployment of security updates to mitigate known vulnerabilities.
6) *Risk Management*: Covers organizational strategies, policies, and procedures aimed at minimizing cybersecurity risks.
7) *Physical Security*: Involves the protection of physical infrastructure and environments hosting critical systems and data.

This classification supports a structured and interpretable validation process, enabling the evaluation of semantic alignment between best practices and system-generated CWE assignments.

A. Result Analysis

To evaluate the effectiveness of the proposed semantic alignment methodology, a classification-based validation approach was adopted. Each CWE assigned to a best practice was evaluated based on its categorical alignment with a predefined taxonomy of cybersecurity domains. The assignment was considered correct if at least one of the categories associated with the identified CWE matched any of the categories assigned to the corresponding best practice. Conversely, if no categorical overlap was found, the assignment was classified as incorrect.

TABLE I
Accuracy results for single-model vs. three-model ensemble configurations

Configuration Name	Accuracy (Ratio)	Accuracy (%)
Single baseline model	62/125	49.6%
3 Baseline Models	80/125	64%

Formally, the overall performance was measured using the accuracy metric, defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Assignments}}{\text{Total Number of Assignments}} \tag{2}$$

Table I compares the performance of two configurations: one using a single baseline model and the other employing an ensemble of three models through a voting mechanism. The ensemble-based configuration achieved an accuracy of 64%, representing a substantial improvement of approximately 15 percentage points over the single-model configuration, which achieved 49.6%.

These results underscore the benefits of model diversity and ensemble learning in enhancing the quality of semantic similarity assessments. By integrating multiple models with distinct characteristics, the system can compensate for individual limitations and produce more robust and reliable risk score associations.

V. Conclusion

While the proposed architecture offers a promising and fully automated methodology for assessing misconfiguration risks in OT environments, several limitations merit critical attention. By leveraging semantic similarity via sentence embeddings and systematically mapping best practices to the CWE and CVSS frameworks, the system introduces an innovative approach to bridging the gap between high-level security guidelines and quantitative risk metrics. However, the absence of a dedicated benchmark dataset for validating the alignment between best practices and CWE entries introduces an inherent limitation to the evaluation's rigor.

Although the STS Benchmark serves as a reasonable proxy for assessing model reliability, it does not capture the full complexity and domain specificity of cybersecurity language. Moreover, the achieved accuracy of 64% while notably outperforming single-model baselines still leaves a considerable margin for misclassification, which may be critical in high-stakes industrial contexts. A further limitation lies in the sole reliance on semantic similarity measures, which may fail to capture contextual dependencies or domain-specific nuances embedded in OT configurations. The lack of human-in-the-loop validation or expert feedback mechanisms also constrains the system's capacity to refine its associations and improve interpretability.

Future research should consider the integration of supervised learning techniques, supported by the development of annotated datasets tailored to misconfiguration scenarios. Enhancing the model with contextual metadata, such as system architecture or deployment parameters, could also improve the

accuracy and relevance of risk scores. Furthermore, efforts should be directed toward improving the real-time adaptability and explainability of model decisions to support deployment in operational environments.

In conclusion, the framework presented in this work establishes a solid foundation for automated misconfiguration risk assessment in OT systems. Nevertheless, addressing its current limitations is essential to ensure the robustness, precision, and practical applicability of the methodology in real-world industrial cybersecurity scenarios.

REFERENCES

[1] G. Assenza, L. Faramondi, G. Oliva, and R. Setola, "Cyber threats for operational technologies," *International Journal of System of Systems Engineering*, vol. 10, no. 2, pp. 128–142, 2020.

[2] F. Fotis, "Economic impact of cyber attacks and effective cyber risk management strategies: A light literature review and case study analysis," *Procedia Computer Science*, vol. 251, pp. 471–478, 2024.

[3] S. Christey, J. Kenderdine, J. Mazella, and B. Miles, "Common weakness enumeration," *Mitre Corporation*, 2013.

[4] C. Vulnerabilities, "Common vulnerabilities and exposures," *The MITRE Corporation,[online] Available: https://cve. mitre. org/index. html*, 2005.

[5] S. Loureiro, "Security misconfigurations and how to prevent them," *Network Security*, vol. 2021, no. 5, pp. 13–16, 2021.

[6] K. Scarfone and P. Mell, "The common configuration scoring system (ccss): Metrics for software security configuration vulnerabilities," *NIST interagency report*, vol. 7502, 2010.

[7] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[8] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1–14. [Online]. Available: https://aclanthology.org/S17-2001

[9] Department of Homeland Security (DHS)'s National Cybersecurity, "Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies," https://www.cisa.gov/sites/default/files/recommended_practices/NCCIC_ICS-CERT_Defense_in_Depth_2016_S508C.pdf, 2016, online; accessed 8 May 2025.