Integrating LLMs and STIX with Case-Based Reasoning for Enhanced Cyber Threat Analysis

Marc Krüger[1,2,*,1]

[1]Stiftung Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

Abstract

The integration of Large Language Models (LLMs) such as ModernBERT, GPT-3.5, and LLaMA 3.2 into cybersecurity has opened new possibilities for analyzing unstructured threat intelligence. Simultaneously, structured standards like STIX (Structured Threat Information Expression) enable interoperable and machine-readable threat descriptions. This paper explores the intersection of LLMs and STIX to assess how semantic analysis through LLMs can enhance the quality and utility of structured threat intelligence. It further investigates how combining unstructured natural language insights with structured formats supports more efficient, scalable, and actionable cybersecurity processes.

Keywords

Cyber Threat Intelligence (CTI), STIX, Large Language Models (LLMs), Natural Language Processing (NLP), Threat Detection, GPT-3.5, ModernBERT, LLaMA, Structured Data, Semantic Analysis

1. Introduction

The growing volume and sophistication of cyber threats call for more adaptive and intelligent analysis frameworks [1]. Conventional rule-based detection methods frequently fall short in effectively identifying new and evolving threat vectors [2]. As threat actors continuously evolve their tactics, techniques, and procedures (TTPs), cybersecurity systems must integrate both structured and unstructured data sources to provide comprehensive insights. Automated frameworks like TTPXHunter demonstrate how unstructured CTI reports can be transformed into structured, STIX-compatible TTP representations [3], while tools such as TTPDrill show the feasibility of mapping semantic threat actions to STIX entities [4]. This paper explores how modern NLP technologies, specifically Large Language Models, can be combined with structured standards like STIX to create a synergistic approach to threat intelligence.

1.1. Problem Identification

Cyber threat intelligence is frequently scattered across various formats, ranging from unstructured text in reports to highly formalized threat indicators in standardized formats like STIX [5, 6]. The disparity between human-readable and machine-readable intelligence hinders automated analysis and operationalization of threat data. Additionally, existing analytical models

Workshop SIG Knowledge Management (FG WM) at KI 2025 September 16, 2025, Potsdam, Germany.

krueger.hannover@t-online.de (M. Krüger)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-06-07

often lack the semantic depth required to interpret context-rich language used in threat descriptions. There is a gap in aligning natural language processing with formal threat modeling standards in a scalable and interpretable manner.

Illustrative Example

To illustrate this challenge, consider a ransomware campaign reported in a threat bulletin. In its unstructured form, the report might describe that "a healthcare provider was targeted with phishing emails containing malicious attachments, which, once opened, deployed a ransomware variant that encrypted patient data and demanded payment in Bitcoin." While human analysts can immediately recognize this as a ransomware incident, automated processing remains difficult.

When expressed in STIX, the same information can be structured into machine-readable entities such as:

- `malware` object: *name = "RansomwareX", type = ransomware*
- `attack-pattern` object: *name = "Phishing", associated tactic = "Initial Access"*
- `indicator` object: *observable = "malicious attachment hash", related domain = "suspicious sender"*
- `identity` object: *target = healthcare organization*

This structured representation ensures interoperability across systems but cannot capture the full semantic richness of threat narratives. Leveraging LLMs to interpret unstructured text and align it with STIX entities enables richer, reusable knowledge for more effective cyber threat analysis.

1.2. Research Objectives

This paper aims to investigate how the integration of LLMs with the STIX framework can improve the processing and interpretation of cyber threat intelligence. The key objectives include:

- To assess the capacity of LLMs to extract, classify, and enrich threat indicators from unstructured sources.
- To evaluate the semantic alignment between LLM-generated insights and the structured representations used in STIX.
- To propose a methodological approach for combining NLP outputs with STIX-based threat modeling for enhanced detection, correlation, and decision-making.

1.3. Research Question

Based on the identified gap between unstructured natural language threat reports and structured STIX-based intelligence, this paper addresses the following research question:

> *How can Large Language Models (LLMs) be effectively integrated with the STIX framework within a Case-Based Reasoning (CBR) paradigm to improve the semantic interpretation, interoperability, and reuse of cyber threat intelligence?*

1.4. Theoretical and Practical Implications

From a theoretical perspective, this research contributes to bridging the gap between natural language processing and formalized threat intelligence standards. It extends prior work on CBR and semantic similarity by aligning LLM-based representations with structured CTI models. From a practical perspective, the proposed framework provides cybersecurity analysts with an explainable, reusable, and interoperable method for case comparison and enrichment. This supports more efficient detection, correlation, and decision-making in operational environments where rapid response to evolving threats is crucial.

2. Related Work

The integration of structured cyber threat intelligence with advanced natural language processing methods has gained increasing attention in recent years [7]. Structured formats such as STIX have become essential for enabling interoperability in cyber defense systems by providing a common language for describing threat indicators, incidents, and adversary behavior patterns [8]. Early research efforts primarily focused on the creation and dissemination of STIX-formatted data using platforms such as TAXII, without addressing the semantic interpretation of unstructured sources [9].

Recent studies have examined the use of machine learning and NLP for enriching threat intelligence. For example, Li et al.[10] presented a case-based reasoning (CBR) [11][12] framework that leverages ontologies to enhance STIX-based threat representation and matching. Their work highlights the importance of combining formal threat modeling with similarity-based reasoning to improve detection accuracy.

In a related domain, Zaw and Vasupongayya (2019) proposed a CBR-based adaptation mechanism for phishing detection, demonstrating how textual indicators can be reused to improve classification performance in mobile environments [13]. Although not directly focused on STIX, their approach underlines the potential of CBR in handling dynamic threat contexts [14].

Large Language Models (LLMs) such as ModernBERT and GPT-3.5 have been increasingly employed for cyber threat detection and classification tasks [15]. Their ability to process large volumes of unstructured threat reports and extract contextual indicators has shown promise in narrowing the gap between human-authored threat narratives and machine-readable representations [16]. However, most existing LLM-based studies operate independently of formal CTI frameworks like STIX, limiting their practical interoperability [17].

While some experimental systems attempt to bridge the gap between LLMs and structured formats, these efforts often remain domain-specific or lack scalability [18]. To date, a comprehensive methodology that aligns the semantic capabilities of LLMs with the rigor of STIX-based modeling in a reusable and explainable framework remains underexplored [19].

For example, Xu et al. (2024) present IntelEX, a framework that leverages LLMs to extract attack-level CTI in STIX-compatible structure, but it focuses primarily on Advanced Persistent Threats and does not integrate CBR for reusable case evaluation [18]. Similarly, Zhang et al. (2024) propose AttacKG+, a method to construct comprehensive attack knowledge graphs from CTI, yet it omits the implementation of a formal CBR loop or explainable STIX integration [20].

This paper contributes to the field by proposing an integrated approach that utilizes LLMs for semantic enrichment and CBR for similarity-based case evaluation, all mapped onto a STIX-compatible structure.

3. Methodology

The system is based on a hybrid architecture that combines structured threat modeling (STIX), semantic similarity assessment using Large Language Models (LLMs), and case-based retrieval mechanisms. The overall goal is to assess and retrieve cyber threat cases that are semantically related, based on both textual descriptions and formal STIX object references. The overall workflow of the proposed system is illustrated in Figure 1, which highlights the integration of unstructured input, semantic enrichment via LLMs, structured mapping to STIX entities, and subsequent reasoning through CBR. This study's methodology is based on established methods in case-based reasoning (CBR) and semantic similarity analysis. Following the CBR cycle as defined by Aamodt and Plaza [11], the system implements case representation, retrieval, and adaptation mechanisms tailored to cyber threat intelligence (CTI). For semantic similarity computation, the study adopts embedding-based comparison methods as proposed by Reimers and Gurevych [21], which have become a standard in evaluating textual relatedness. Finally, the enriched and retrieved cases are aligned with the Structured Threat Information Expression (STIX) 2.1 standard [**?**], ensuring interoperability and structured representation within established CTI frameworks. The choice of models in this study was guided by their complementary strengths. ModernBERT represents a state-of-the-art embedding model optimized for semantic similarity tasks, offering efficient vector representations well-suited for document comparison in CTI [22, 23]. GPT-3.5-Turbo-Instruct was included as an instruction-tuned model with strong generalization capabilities, allowing the evaluation of how conversational LLMs handle CTI similarity judgments [24, 25]. LLaMA 3.2, as a recent open-source foundation model, combines high performance with reproducibility and transparency, making it attractive for research contexts where local deployment and explainability are critical [26]. Together, these models cover a spectrum from embedding-focused architectures to instruction-following and open-source general-purpose LLMs, providing a robust basis for evaluating semantic similarity in cyber threat intelligence.

3.1. System Architecture

The technical implementation relies on a three-tier architecture:

- **Backend:** MySQL Community Server (v8.0) is used for persistent storage of cases, STIX objects, and evaluation results.
- **Application Layer:** The Java Development Kit (JDK 21) and Apache Tomcat 9 serve as the runtime environment for backend logic and servlet handling.
- **Frontend:** An interactive web interface allows the input and management of reference documents and evaluation cases.

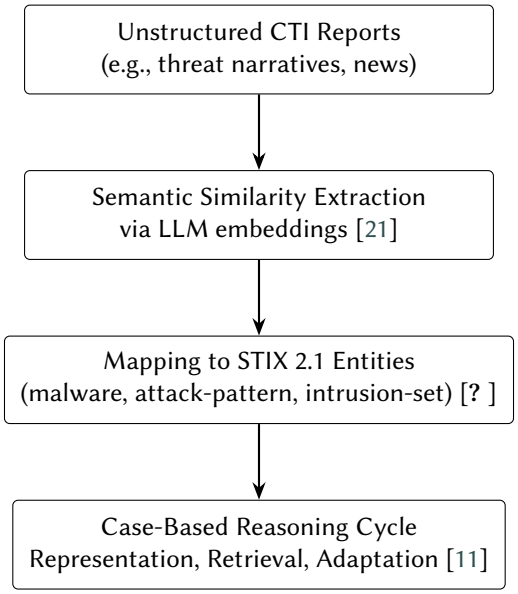

Figure 1: Methodological workflow integrating LLM-based semantic similarity, STIX-based structured intelligence, and CBR mechanisms.

3.2. LLM Evaluation and Metric Computation

Each pair of documents is automatically analyzed using multiple large language models (LLMs), such as ModernBERT, GPT-3.5-Turbo, and LLaMA 3.2. The model output is interpreted as `similar` or `not similar` based on predefined similarity thresholds. This approach follows established practices in semantic similarity research, such as ModernBERT's embedding comparison with cosine similarity [21], and the threshold-based classification paradigm demonstrated in ModernBERT for determining pairwise relevance [27].

The model predictions are then compared to the manually defined Ground Truth labels. Based on this comparison, the following evaluation metrics are computed:

- **True Positive (TP):** Model and evaluator both label a pair as similar.
- **True Negative (TN):** Model and evaluator both label a pair as not similar.
- **False Positive (FP):** Model classifies a pair as similar, while the evaluator labels it as not similar.
- **False Negative (FN):** Model classifies a pair as not similar, while the evaluator identifies it as similar.

From these values, standard performance metrics are calculated as follows:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics allow for a quantitative comparison of the semantic similarity performance of each LLM and help assess their applicability in the context of structured threat intelligence mapping.

3.3. Mapping to STIX Entities

The enriched information is mapped to STIX (Structured Threat Information Expression) objects. Entities such as `malware`, `attack-pattern`, `intrusion-set`, and `indicator` are created or linked based on the semantics of the extracted and enriched data. This structured representation ensures interoperability with other CTI systems and enables downstream automation, such as correlation and alert generation.

3.4. Case Storage and Similarity Assessment

Each enriched and structured case is stored in a case base using a Case-Based Reasoning (CBR) model. When a new case is introduced, it is compared to existing cases using a multi-dimensional similarity function. This function considers both lexical overlap (e.g., cosine similarity of term vectors) and semantic similarity (e.g., LLM-based embeddings).

3.5. Retrieval and Adaptation

The most similar cases are retrieved to support decision-making. Adaptation mechanisms refine the retrieved solutions to better fit the specifics of the current case. The resulting suggestions support analysts by offering evidence-based recommendations derived from prior experiences, while remaining transparent and explainable.

3.6. Evaluation Setup

To evaluate the proposed system, a ground truth dataset was created through expert annotation of semantically similar and dissimilar case pairs. The case base comprised **29 individual cases**, from which **48 annotated pairwise comparisons** were selected for evaluation. Performance was measured using standard metrics such as precision, recall, and F1-score. Additionally, the results were compared to baseline methods using only LLMs or traditional keyword-matching to assess the added value of the integrated framework.

The methodological framework follows a sequential workflow that begins with the ingestion of raw threat reports and proceeds through a series of refinement stages. Initially, relevant keywords are automatically extracted using STIX, which isolates key phrases and terms from the unstructured text and showed STIX-objects for the new case [28].

Subsequently, large-scale language models generate semantic representations for the extracted keywords, enabling the inference of higher-level threat indicators. These indicators are then

aligned with known cyber threat concepts as defined in the STIX standard, such as malware families, attack techniques, and actor profiles.

The semantically enriched cases are stored in a structured case base, where each entry includes STIX-compatible elements. When a new case is submitted, it undergoes the same preprocessing and enrichment pipeline before being compared to existing cases through a similarity analysis. The process ensures that historical knowledge is reused for decision support while maintaining explainability through the explicit mapping to structured concepts.

This integrated flow supports transparent, traceable, and adaptive threat analysis by fusing data-driven models with formal threat representations.

4. Evaluation

To assess the effectiveness of the proposed framework, a multi-layered evaluation was conducted using both qualitative expert annotations and quantitative performance metrics. The aim was to determine whether the integration of LLMs and STIX within a CBR framework improves the identification and retrieval of semantically similar threat cases.

4.1. Ground Truth Construction

To support the evaluation of semantic similarity outputs generated by large language models (LLMs), a Ground Truth dataset was systematically constructed through manual annotation. This process involved the selection of document pairs within the cyber threat intelligence (CTI) domain, each annotated with binary similarity labels (`similar` / `not similar`) based on qualitative expert assessment.

Annotation decisions were guided by a set of well-defined evaluation criteria, summarized in Table 1. These criteria were designed to capture semantic alignment beyond lexical overlap, incorporating higher-level attributes such as described attack techniques, affected targets, adversary profiles, and contextual depth.

An illustrative example of Ground Truth annotation is provided below:

- **Document A:** A technical intelligence report describing a ransomware variant targeting hospital IT systems through phishing emails, resulting in encrypted patient data.
- **Document B:** A cybercrime case summary detailing a ransomware intrusion into a healthcare facility, executed via credential phishing and followed by file encryption.

Assigned Ground Truth Label: `similar`
Annotation Justification:

- *Topic:* Both pertain to ransomware.
- *Target:* Healthcare organizations.
- *Technique:* Use of phishing and subsequent encryption of systems.
- *Attacker Profile:* Implicit alignment based on technique and target.
- *Distribution Strategy:* Mass email phishing vectors.
- *Depth of Description:* Technical and detailed in both cases.

Table 1
Qualitative criteria for ground-truth annotation (inspired by SemEval-style semantic similarity benchmarks [29])

Criterion	Classification	Rationale
Topic	Similar / Not Similar	Assess whether both documents pertain to the same overall threat type (e.g., ransomware, phishing).
Target	Similar / Not Similar	Determine if the entities affected by the attacks are of comparable type (e.g., critical infrastructure, individuals).
Technique	Similar / Not Similar	Evaluate the similarity of employed tactics, techniques, and procedures (TTPs), aligned to MITRE ATT&CK when feasible.
Attacker Profile	Similar / Not Similar	Consider whether the behavioral characteristics or inferred identity of the adversary are described in both cases.
Distribution Strategy	Similar / Not Similar	Identify overlap in how the attack was propagated (e.g., spear-phishing, watering hole attacks).
Depth of Description	Similar / Not Similar	Examine the richness and detail level of the reports to ensure comparable granularity.
Case Type	Similar / Not Similar	Determine whether both reports describe distinct incidents or refer to the same campaign or malware family.

- *Case Type:* Distinct incidents within the same malware family.

This annotated label serves as the reference classification against which LLM-generated similarity predictions are benchmarked during evaluation.

4.2. Baseline Comparison

To validate the added value of the integrated system, its performance was compared to several baseline approaches:

- A keyword-only approach using cosine similarity over term frequency-inverse document frequency (TF-IDF) vectors.
- A pure LLM-based approach using sentence embeddings generated by ModernBERT and GPT-3.5.
- A CBR system using manually defined rules without semantic enrichment.

4.3. Results Summary

The integrated framework outperformed baseline approaches in retrieving relevant threat cases. Notably, incorporating STIX entities enhanced interpretability, while semantic enrichment via LLMs improved retrieval precision. These results demonstrate that combining structured CTI with LLMs and case-based adaptation leads to a more robust and explainable solution.

5. Results and Discussion

The updated evaluation confirms the effectiveness of integrating structured threat modeling (STIX) with LLM-based semantic embeddings in a Case-Based Reasoning framework.

Table 2 reports extended performance metrics derived from the confusion matrix. Among the evaluated models, **LLaMA 3.2** achieved the highest accuracy (91.67%) and F1-score (94.87%), indicating both precise and consistent performance. **GPT-3.5-Turbo-Instruct** delivered perfect precision (100%) but lower recall (76.32%), which reduced its overall F1-score. In contrast, **ModernBERT** achieved perfect recall (100%) but showed the lowest accuracy (79.17%) due to a higher number of false positives, which limited its overall effectiveness.

Table 2
Extended performance metrics per algorithm

Algorithm	TP	TN	FP	FN	Accuracy	Precision	Recall	F1-Score
ModernBERT	38	0	10	0	79.17%	79.17%	100.00%	88.37%
GPT-3.5-Turbo-Instruct	29	10	0	9	81.25%	100.00%	76.32%	86.57%
LLaMA 3.2	37	7	3	1	91.67%	92.50%	97.37%	94.87%

Table 3 provides a concise overview of the overall prediction accuracy across **48 case comparisons**.

Table 3
Overall accuracy comparison

Algorithm	Correct Predictions	Total Evaluations	Accuracy (%)
ModernBERT	38	48	79.17%
GPT-3.5-Turbo-Instruct	39	48	81.25%
LLaMA 3.2	44	48	91.67%

LLaMA 3.2 correctly classified **44 out of 48 case comparisons**, reflecting strong generalizability and balanced performance.

The results indicate that instruction-tuned LLMs such as GPT-3.5 and LLaMA 3.2 significantly surpass older models in terms of both predictive accuracy and operational stability. Moreover, integrating semantic similarity with structured STIX object references enhances both traceability and explainability.

While **ModernBERT** achieved high recall, its elevated false-positive rate limited its overall effectiveness. **GPT-3.5-Turbo-Instruct**, by contrast, maintained perfect precision but at the expense of overlooking some relevant cases. **LLaMA 3.2** emerged as the most balanced and deployment-ready model for cyber threat intelligence applications.

Comparison with Related Work

Li et al. [10] introduced AttacKG, a pipeline for constructing technique knowledge graphs (TKGs) from CTI reports. Their approach achieved strong results in entity and dependency extraction (F1=0.887/0.896) and technique identification (F1=0.789), but remained limited to extraction and aggregation. The framework presented in this study extends beyond these capabilities

by integrating semantic similarity assessment via LLM embeddings with STIX mappings and CBR-based case reuse, reaching an F1-score of 94.87% with LLaMA 3.2 on pairwise similarity tasks.

Xu et al. [18] proposed IntelEX, an LLM-based system that identifies logical TTP sequences and contextual insights. It achieved F1=0.792 for technique identification across 1,769 reports and demonstrated utility in downstream tasks such as Sigma rule generation (F1=0.929). In contrast, the present framework emphasizes semantic similarity and retrieval of related cases, offering explainable and reusable intelligence rather than solely extraction or rule generation.

Zhang et al. [20] presented AttacKG+, a fully automatic LLM-based method for constructing temporal attack knowledge graphs with modules for rewriting, parsing, identification, and summarization. While effective for temporal modeling and attack reconstruction, AttacKG+ does not incorporate case similarity evaluation or reasoning. The framework analyzed in this paper addresses this gap by combining semantic enrichment, STIX alignment, and CBR, yielding both higher retrieval accuracy (accuracy 91.67%, F1=94.87%) and greater interpretability.

Overall, the contribution of this study lies in extending beyond extraction and graph construction towards similarity-driven case reuse and decision support.

6. Conclusion and Future Work

This study demonstrates that the integration of Large Language Models (LLMs), structured threat intelligence via STIX, and Case-Based Reasoning (CBR) can significantly improve the quality and explainability of cyber threat analysis. The evaluated system combines semantic understanding of unstructured threat descriptions with structured mappings to STIX entities, enabling both high accuracy and traceability.

Among the tested approaches, **LLaMA 3.2** achieved the best overall performance with the highest accuracy (91.67%) and F1-score (94.87%), correctly classifying **44 out of 48 case comparisons**. **GPT-3.5-Turbo-Instruct** demonstrated perfect precision (100%), but its lower recall (76.32%) led to fewer correctly identified cases. **ModernBERT**, while detecting all relevant cases with perfect recall (100%), produced a higher number of false positives, resulting in the weakest overall accuracy (79.17%). These results confirm the value of LLM-assisted reasoning frameworks for practical threat classification and reuse of historical cases. Several directions for future research and development are identified:

- **Integration of domain-adapted LLMs:** Fine-tuning LLMs on cybersecurity-specific corpora (e.g., MITRE reports, CVEs, threat bulletins) could improve contextual sensitivity and reduce false negatives.
- **Multilingual and cross-domain capability:** Expanding the system to support multiple languages and applying it to cross-sectoral domains such as healthcare or critical infrastructure could broaden its applicability.
- **Explainability and visualization:** Enhancing the interpretability of LLM outputs through visualization of token attention or STIX-linked semantic paths may improve analyst trust and adoption.
- **Integration with real-time feeds:** Incorporating live data streams (e.g., TAXII, SIEM logs) would enable the system to dynamically adapt to emerging threats and continuously

enrich the case base.

In conclusion, the proposed framework provides a robust foundation for future developments in intelligent, transparent, and scalable cyber threat analysis. Combining structured knowledge representations with the power of language models offers promising pathways for decision support and threat detection.

Data Availability

The dataset underlying this study was compiled from publicly accessible newspaper articles and online portals reporting on cybercrime cases between 2010 and 2025 and consists of **29 individual cases** and **48 annotated pairwise comparisons**. The dataset supporting this study is openly available on Zenodo at doi:10.5281/zenodo.16985159 [30]. Since the original URLs are no longer available, only the structured case data (e.g., actors, methods, affected domains, and summaries) are published. The dataset has been anonymized where necessary and is released under a CC-BY 4.0 license to ensure reproducibility and reusability for the research community.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Fujii, N. Kawaguchi, T. Shigemoto, T. Yamauchi, Extracting and analyzing cybersecurity named entity and its relationship with noncontextual IOCs from unstructured text of CTI sources, Journal of Information Processing 31 (2023) 578–590.

[2] W. Alasmary, U. Mbanaso, G. Epiphaniou, A deep learning approach for cyber threat intelligence, Computers & Security 104 (2021) 102201.

[3] N. Rani, B. Saha, V. Maurya, S. K. Shukla, Ttpxhunter: Actionable threat intelligence extraction as TTPs from finished cyber threat reports, arXiv preprint arXiv:2403.03267 (2024). Includes STIX-compatible TTP mapping.

[4] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, X. Niu, Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources, in: ACSAC '17: Annual Computer Security Applications Conference, Orlando, FL, USA, 2017. doi:10.1145/3134600.3134646.

[5] G. Siracusano, D. Sanvito, R. Gonzalez, et al., Time for aCTIon: Automated analysis of cyber threat intelligence in the wild, arXiv preprint arXiv:2307.10214 (2023). URL: https://arxiv.org/abs/2307.10214.

[6] J. Wheelus, Others, Probabilistic measurement of CTI quality for large numbers of structured and unstructured threat reports, Electronics (MDPI) 14 (2022) 1826. doi:10.3390/electronics14091826.

[7] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, R. R. KA, M. Conti, et al., NLP-based techniques for cyber threat intelligence, Computer Science Review 58 (2025) 100765.

[8] OASIS Cyber Threat Intelligence (CTI) Technical Committee, STIX 2.1 specification, 2020. URL: https://docs.oasis-open.org/cti/stix/v2.1/cs01/stix-v2.1-cs01.html, version 2.1 of the STIX standard defines domain objects such as indicators, TTPs, campaigns, and more.

[9] A. Khalid, STIX/TAXII: All Your Questions Answered, Blog, Anomali, 2017. URL: https://www.anomali.com/blog/stix-taxii-all-your-questions-answered, describes STIX as the 'what' and TAXII as the 'how' in CTI sharing.

[10] Z. Li, J. Zeng, Y. Chen, Attackg: Constructing technique knowledge graph from cyber threat intelligence reports, arXiv preprint arXiv:2111.07093 (2021). URL: https://arxiv.org/abs/2111.07093.

[11] A. Aamodt, E. Plaza, Case-Based reasoning: foundational issues, methodological variations, and system approaches, AI communications 7 (1994) 39–59. URL: https://doi.org/10.3233/aic-1994-7104. doi:10.3233/aic-1994-7104.

[12] K.-D. Althoff, Case-based reasoning, in: Handbook of Software Engineering and Knowledge Engineering: Volume I: Fundamentals, World Scientific, 2001, pp. 549–587.

[13] S. K. Zaw, S. Vasupongayya, A case-based reasoning approach for automatic adaptation of classifiers in mobile phishing detection, Journal of Computer Networks and Communications 2019 (2019) 1–14. doi:10.1155/2019/7198435.

[14] H. Kim, L. Chen, AI-driven case-based reasoning for real-time threat detection in cybersecurity, Expert Systems with Applications 213 (2023) 119002.

[15] A. Alsabban, I. Ahmad, M. A. Alzain, A survey on cyber threat intelligence: Technologies, applications and research challenges, Journal of Information Security and Applications 73 (2023) 103530.

[16] E. Mezzi, F. Massacci, K. Tuma, Large language models are unreliable for cyber threat intelligence, arXiv preprint arXiv:2503.23175 (2025). URL: https://arxiv.org/abs/2503.23175.

[17] A. Formato, From unstructured threat intelligence to STIX 2.1 bundles with generative AI, Medium Blog, 2024. URL: https://medium.com/@antonio.formato/from-unstructured-threat-intelligence-to-stix-2-1-bundles-with-generative-ai-1065ce399e63.

[18] M. Xu, H. Wang, J. e. a. Liu, Intelex: A LLM-driven attack-level threat intelligence extraction framework, arXiv preprint arXiv:2412.10872 (2024). URL: https://arxiv.org/abs/2412.10872.

[19] M. T. Alam, D. Bhusal, L. Nguyen, N. Rastogi, Ctibench: A benchmark for evaluating llms in cyber threat intelligence, Advances in Neural Information Processing Systems 37 (2024) 50805–50825.

[20] Y. Zhang, T. Du, Y. e. a. Ma, Attackg+: Boosting attack knowledge graph construction with large language models, arXiv preprint arXiv:2405.04753 (2024).

[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019. URL: https://arxiv.org/abs/1908.10084.

[22] Y. Yamagishi, T. Kikuchi, S. Hanaoka, T. Yoshikawa, O. Abe, ModernBERT is more efficient than conventional BERT for chest CT findings classification in japanese radiology reports, arXiv preprint arXiv:2503.05060 (2025).

[23] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024).

[24] N. Koneva, A. L. G. Navarro, A. Sánchez-Macián, J. A. Hernández, M. Zukerman, Ó. G. de Dios, Introducing large language models as the next challenging internet traffic source, arXiv preprint arXiv:2504.10688 (2025).
[25] Y. Fengrui, Y. Du, Few-shot learning of TTPs classification using large language models, 2024.
[26] I. Azaiz, N. Kiesler, S. Strickroth, A. Zhang, Open, small, rigmarole–evaluating Llama 3.2 3b's feedback for programming exercises, arXiv preprint arXiv:2504.01054 (2025).
[27] Z. Xu, N. Cristianini, QBERT: Generalist model for processing questions, 2022. URL: https://arxiv.org/abs/2212.01967. arXiv:2212.01967.
[28] T. Li, J. Zhang, Y. Wang, An ontology-based CBR framework for cyber threat intelligence sharing using Fuji, in: Proceedings of the 17th International Conference on Availability, Reliability and Security, 2022, pp. 1–8.
[29] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez Agirre, R. Mihalcea, G. Rigau Claramunt, J. Wiebe, Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation (2016).
[30] M. Krüger, Cybercrime case dataset: 29 individual cases and 48 pairwise comparisons (2010–2025), 2025. URL: https://doi.org/10.5281/zenodo.16985159. doi:10.5281/zenodo.16985159, dataset licensed under CC-BY 4.0. All copyrights of the original case descriptions remain with the respective publishers and portals.