Future Generation Computer Systems 173 (2025) 107877

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Can LLM-generated misinformation be detected: A study on Cyber Threat Intelligence

He Huang [a], Nan Sun [a,*], Massimiliano Tani [a], Yu Zhang [a], Jiaojiao Jiang [b], Sanjay Jha [b]

[a] University of New South Wales, Northcott Dr, Campbell, Canberra, 2600, ACT, Australia
[b] University of New South Wales, High St, Kensington, Sydney, 2052, NSW, Australia

ARTICLE INFO

Keywords:
Cyber security
Artificial intelligence
Human-centric

ABSTRACT

Given the increasing number and severity of cyber attacks, there has been a surge in cybersecurity information across various mediums such as posts, news articles, reports, and other resources. Cyber Threat Intelligence (CTI) involves processing data from these cybersecurity sources, enabling professionals and organizations to gain valuable insights. However, with the rapid dissemination of cybersecurity information, the inclusion of fake CTI can lead to severe consequences, including data poisoning attacks. To address this challenge, we have implemented a three-step strategy: generating synthetic CTI, evaluating the quality of the generated CTI, and detecting fake CTI. Unlike other subdomains, such as fake COVID news detection, there is currently no publicly available dataset specifically tailored for fake CTI detection research. To address this gap, we first establish a reliable groundtruth dataset by utilizing domain-specific cybersecurity data to fine-tune a Large Language Model (LLM) for synthetic CTI generation. We then employ crowdsourcing techniques and advanced synthetic data verification methods to evaluate the quality of the generated dataset, introducing a novel evaluation methodology that combines quantitative and qualitative approaches. Our comprehensive evaluation reveals that the generated CTI cannot be distinguished from genuine CTI by human annotators, regardless of their computer science background, demonstrating the effectiveness of our generation approach. We benchmark various misinformation detection techniques against our groundtruth dataset to establish baseline performance metrics for identifying fake CTI. By leveraging existing techniques and adapting them to the context of fake CTI detection, we provide a foundation for future research in this critical field. To facilitate further research, we make our code, dataset, and experimental results publicly available on GitHub.

1. Introduction

Currently, Internet users and organizations face numerous challenges when it comes to protecting against cyber attacks, including their recent general increase and the seeming persistence of devious threat actors. Cyber Threat Intelligence (CTI) refers to "knowledge, skills, and experience-based information concerning the occurrence and assessment of both cyber and physical threats that are intended to help mitigate potential attacks and harmful events occurring in cyberspace", which can help organizations stay ahead of the ever-changing threat landscape [1,2]. The process of analyzing unstructured data on cyber threats involves transforming it into structured information that includes various aspects such as background, mechanism, indicators of compromise, impact, and actionable recommendations of potential threats. The structured information can then be used to provide relevant insights and guidance for decision-making in different areas in the form of strategic, operational, tactical, and technical actions. Furthermore, organizations can use detailed information on current and emerging threats to make informed decisions on various aspects of cybersecurity in the fields of intrusion detection, real-time analytics, forensic investigation, and threat hunting. Since a significant portion of data in cybersecurity is in the form of written words (i.e., text), extracting valuable insights from this rich data source is crucial in the analysis of CTI. Through understanding and utilizing CTI effectively, organizations can enhance their cyber resilience and proactively prevent cyber attacks.

With the growing volume and diversity of cybersecurity data and the advancements in generative Artificial Intelligence (AI), there is a concern that malicious individuals may produce and circulate fake CTI samples. These fake samples could potentially harm security systems by conducting data poisoning attacks, generating incorrect security alerts, and compromising AI-based cyber defence models. For example, the

* Corresponding author.
E-mail address: nan.sun@unsw.edu.au (N. Sun).

https://doi.org/10.1016/j.future.2025.107877
Received 4 November 2024; Received in revised form 6 March 2025; Accepted 22 April 2025
Available online 8 May 2025

dissemination of misleading threat intelligence reports during high-profile cyber incidents has resulted in incorrect attribution of attacks, leading to geopolitical conflicts and delays in response actions [3–5]. In November 2021, unidentified hackers used the FBI mail server to send spam on a large scale and claimed that a cyber attack was taking place by publishing fake CTIs on open source platforms [6]. Similarly, fake CTI has been observed in cyber-espionage activities, where fabricated CTI reports have been intentionally planted to divert attention away from the true perpetrators of an attack. For example, in February 2023, there was a CTI about the ESXi ransomware attack, claiming that the ESXiArgs ransomware was deployed by exploiting a vulnerability in VMware ESXi. In January 2024, there was also a fake CTI claiming that data from several large organizations such as Procter & Gamble and the City of Toronto was stolen by a zero-day vulnerability in the GoAnywhere platform [7]. As these fake CTIs spread in the open source community, cybersecurity defense systems misuse these fake CTIs as training samples, leading to false positives and omissions of real cyber attacks [8]. These examples demonstrate the tangible risks posed by fake CTI and underscore the importance of developing sophisticated detection mechanisms to maintain the integrity of threat intelligence systems.

Additionally, the strategy of viewing cybersecurity experts as the last line of defense faces challenges. According to recent studies, a significant number of fake CTI samples produced by AI are being incorrectly identified as legitimate by cybersecurity experts and threat hunters [9]. This finding highlights the alarming issue of fake CTI and emphasizes the need for increased awareness and vigilance in addressing this threat. Furthermore, it has been observed that one of the main challenges in utilizing CTI to its full potential is the assurance of its quality and authenticity [1,2]. Therefore, it is crucial to consider the possibility of fake CTI when designing robust cyber defense systems that automatically ingest CTI.

Despite significant advances in CTI analysis, the detection of fake CTI remains a critical challenge, primarily due to the absence of comprehensive public datasets for fake CTI detection. To address this fundamental gap, we propose a novel methodology that leverages the Large Language Model (LLM) to generate synthetic CTI data specifically tailored to the cybersecurity domain. We continuously iterate on the generation and validation process to ensure the quality and authenticity of our fake CTI groundtruth. In this paper, the terms "fake CTI", "synthetic CTI", and "LLM-generated CTI" are used interchangeably to refer to fabricated Cyber Threat Intelligence produced through Large Language Models (LLMs). Furthermore, we define misinformation in the cybersecurity context as incorrect or misleading information regarding cyber threats, which may arise unintentionally (e.g., erroneous threat intelligence reporting) or intentionally (e.g., manipulated or deceptive CTI used for adversarial purposes). In addition, we define "fake news" as fabricated or manipulated information that is intentionally or unintentionally disseminated to mislead readers. In the context of CTI, "fake news" pertains specifically to falsified cybersecurity-related reports or narratives that misrepresent incidents or create deceptive impressions about cyber threats.

Assessing the ability of machine-generated text to mimic human writing remains a significant challenge in natural language processing research [10]. To address this challenge, we develop a novel evaluation framework that combines crowdsourcing methodologies with advanced synthetic data verification techniques. Our framework is designed to determine whether machine-generated content can be indistinguishable from human-written text, focusing on the domain of CTI. By leveraging crowdsourcing, we tap into diverse human perspectives and employ rigorous verification methods to ensure the reliability of our evaluation. The results show that cybersecurity professionals and general participants alike struggle to differentiate LLM-generated CTI from real CTI, achieving only 56.45% accuracy on average. Notably, even IT professionals only reached a 66.22% accuracy rate, demonstrating the effectiveness of our generation approach and highlighting the challenges in distinguishing between synthetic and real CTI samples. Furthermore, our quantitative evaluation confirms that our LLM-generated CTI closely mimics authentic samples across multiple linguistic dimensions, with sentence-level characteristics being nearly indistinguishable from real CTI.

Therefore, we investigate key factors that influence people's ability to distinguish between real and fake CTI, as this can impact the propagation of fake CTI samples. Our analysis reveals that text readability, credibility, and cognitive biases significantly influence human judgments. To address this, we examined the impact of targeted training and observed that cybersecurity experts who underwent structured training improved their detection accuracy to 85.71%. This suggests that systematic exposure to key differentiating features can enhance human judgment. Additionally, we benchmark the detection of synthetic CTI using our groundtruth dataset across three categories of methods: traditional machine learning classifiers, enhanced language models leveraging architectures, and transformer-based approaches. This systematic evaluation provides a foundation for developing more robust misinformation detection systems, specifically for cybersecurity applications.

In summary, our research investigates the detectability of misinformation in CTI generated by fine-tuned LLMs, addressing a critical gap in the cybersecurity domain where no publicly available datasets for fake CTI detection currently exist. We develop a comprehensive methodology by first creating the *first-of-its-kind dataset* using publicly available, expert-validated CTI reports to fine-tune LLMs, enabling the generation of domain-specific synthetic CTI samples. Our approach not only leverages the general capabilities of LLMs but also incorporates *domain expertise through specialized fine-tuning* on authentic CTI reports, ensuring the generated content aligns with the technical and contextual characteristics of real-world CTI. To ensure dataset reliability, we implement a *novel dual validation framework* that combines professional IT expert crowdsourcing with advanced statistical verification methods, providing a robust assessment of the characteristics of authentic versus LLM-generated CTI. Beyond dataset development, we conduct an *in-depth investigation into human factors*, specifically examining how IT expertise and cybersecurity awareness influence the ability to distinguish between authentic and synthetic CTI. This *human-centric analysis* is critical for understanding the propagation and impact of fake CTI in real-world scenarios. Finally, our experiments establish *comprehensive detection benchmarks*, evaluating multiple machine learning classifiers, including traditional and advanced transformer-based models, against human performance. These benchmarks not only identify key factors influencing the accurate identification of synthetic threat intelligence but also provide foundational metrics for future research, making our work a *significant contribution to the growing field of cybersecurity and misinformation detection*.This work presents four major contributions:

- Establishment of the first extensive dataset for fake CTI detection by fine-tuning an LLM on CTI resources verified and annotated by humans, offering a publicly accessible groundtruth dataset.
- Design of a hybrid validation framework that leverages expert crowdsourcing and quantitative metrics to assess the ability of LLM-generated CTI to mimic human-generated content.
- Applying advanced misinformation detection techniques to the benchmark dataset yields initial performance metrics for detecting LLM-generated fake CTI.
- Analysis of critical factors influencing the detection of LLM-generated threat intelligence, providing evidence-based insights to guide future research in securing CTI against misinformation.

The paper is organized as follows: In Section 2, we present the research background of CTI, and the related work in fake CTI detection. We describe our designed methods for fake CTI generation, evaluation, and detection in Section 3. Next, we present our experiments and analyze the results in Section 4. Finally, we conclude and introduce future work in Section 5.2.

2. Background and related work

In this section, we provide an overview of the background of CTI and discuss related research on text generation and misinformation detection.

2.1. Cyber threat intelligence

The CTI lifecycle comprises six phases: direction, collection, processing, analysis, dissemination, and feedback [11]. The collection phase has been the subject of numerous studies, including the iACE method by Liao et al. for automatic discovery of open source cyber threat intelligence (OSCTI) [12] and ThreatRaptor by Gao et al. for log-based cyber threat hunting using OSCTI [13]. Processing phase involves formatting, filtering out redundant information, and making all collected data available to the organization in the form of structured CTI. Barnum et al. proposed Structured Threat Information eXpression (STIX) for standardizing and structuring source information [14], while Husari et al. proposed TTPDrill for automatic extraction of threat behaviors from unstructured text in CTI sources [15]. In addition, Nan et al. [16] propose a novel data-driven approach to rapidly analyze and measure vulnerabilities mentioned on Twitter by integrating intelligence from security experts and social crowds. CTI analytics phase involves transforming processed information into intelligence to inform decision-making, such as using machine learning to detect and predict cyber attacks [17]. Dissemination phase involves getting the structured intelligence output where it needs to go, and sharing is an important means of disseminating CTI [18–21]. Current research primarily focuses on the collection, processing, analysis, and dissemination phases of the CTI lifecycle. However, research on the feedback stage, particularly filtering out fake CTI during processing, is limited.

2.2. CTI datasets and the need for fake CTI detection

Cyber Threat Intelligence (CTI) datasets play a crucial role in cybersecurity research, enabling automated threat detection, real-time monitoring, and decision support. However, existing datasets primarily focus on structuring and summarizing real-world security incidents while largely neglecting the impact of misinformation and synthetically generated CTI [22,23]. The potential risks associated with fake CTI contamination in open-source intelligence platforms such as AlienVault OTX, IBM X-Force Exchange, and Facebook Threat Exchange highlight the need for robust verification mechanisms. Several well-known CTI datasets have been developed to support cybersecurity research, yet they focus primarily on real CTI sources without addressing the risks posed by synthetic or manipulated intelligence. For instance, CASIE [24] is widely used for cyberattack event extraction from news reports but lacks adversarial misinformation samples, making it unsuitable for studying fake CTI. Similarly, the UMBC [25] Cybersecurity Blog Dataset collects cyber threat intelligence from online sources but does not provide structured annotations for identifying fake intelligence. The APT (Advanced Persistent Threat) Notes dataset [26] compiles reports on sophisticated cyber-attacks but does not include adversarially generated misinformation or AI-generated CTI. More recently, CTISum [27] has contributed by summarizing large-scale unstructured CTI data, yet its focus remains on real threats rather than fabricated or misleading reports. Additionally, CTIMiner [28] automates the collection and structuring of CTI from security blogs and malware repositories, yet it does not include a validation mechanism for filtering out misinformation or adversarially generated CTI.

Unlike these existing datasets, our proposed dataset explicitly incorporates both real and synthetic CTI samples, ensuring that AI-driven cybersecurity tools can effectively distinguish between genuine intelligence and fabricated misinformation. By leveraging Large Language Models (LLMs) for CTI generation and implementing a dual-validation framework involving human experts and statistical verification, our dataset provides a unique benchmark for misinformation detection in the cybersecurity domain. This distinction is crucial as misleading CTI can significantly impact security responses, misallocate resources, and introduce vulnerabilities in threat intelligence systems. The inclusion of fake CTI samples enhances the dataset's applicability to training advanced AI models that can improve the resilience of cybersecurity operations against misinformation-based threats.

2.3. Language generation

Text generation is a significant field within Natural Language Processing (NLP) that involves generating human-like language from various forms of input data, such as images, tables, and knowledge bases [29]. It has a range of applications, including machine translation, text summarization, dialogue, and text authoring [30–33]. However, text generation models using mainstream RNN models have slow training speeds. In 2017, Google introduced the transformer structure, which enabled large-scale domain training models. These pre-trained models can learn from unlabeled data, acquiring a vast amount of semantic and syntactic knowledge from the data using unsupervised methods. GPT and BERT are examples of these pre-trained models, and large language models like GPT have surpassed previous models in long text generation [34]. In addition, it is worthwhile to mention that text generation has become a common tool for generating misinformation.

2.4. Misinformation detection in CTI

Misinformation detection has been extensively studied across various domains, ranging from fake news detection to machine-generated text identification. However, limited research has been conducted on misinformation detection within Cyber Threat Intelligence (CTI), where misleading information can directly impact security decision-making. This section reviews existing misinformation detection approaches, emphasizing the unique challenges posed by CTI-specific misinformation and the limitations of prior research.

2.4.1. Human-oriented detection methods

Human-based misinformation detection remains widely used, particularly in cases where text-based misinformation is subtle and context-dependent. Various studies have explored the role of crowdsourced annotation, expert verification, and interactive tools in detecting false information. For example, Dugan et al. [35] designed an interactive website to involve humans in identifying fake information, while Ranade et al. [9] engaged cybersecurity professionals and threat hunters to manually assess the authenticity of CTI samples. These studies show that the human ability to detect misinformation remains unreliable. Research by Rubin et al. [36] suggests that human deception detection accuracy is only slightly above chance, indicating the need for supporting tools in high-stakes contexts. In the CTI domain, Ranade et al. [9] similarly found that cybersecurity professionals struggle to distinguish real CTI from fabricated samples, reinforcing concerns about the effectiveness of human verification alone. To mitigate this challenge, human–computer interaction tools have been proposed to enhance detection capabilities. For example, Gehrmann et al. [37] introduced the Giant Language Model Test Room (GLTR), which visualizes token probability distributions to assist humans in identifying AI-generated text. However, such tools have not been widely tested for CTI-specific misinformation, and their generalizability remains uncertain.

Overall, while human-oriented approaches provide qualitative insights, they face scalability challenges in handling large-scale CTI data. Additionally, human assessments are subject to bias and inconsistencies, particularly in high-stakes cybersecurity contexts where expertise levels vary. Our study builds upon prior research by combining human annotation with statistical and automated verification techniques, ensuring a scalable and reliable CTI misinformation detection framework.

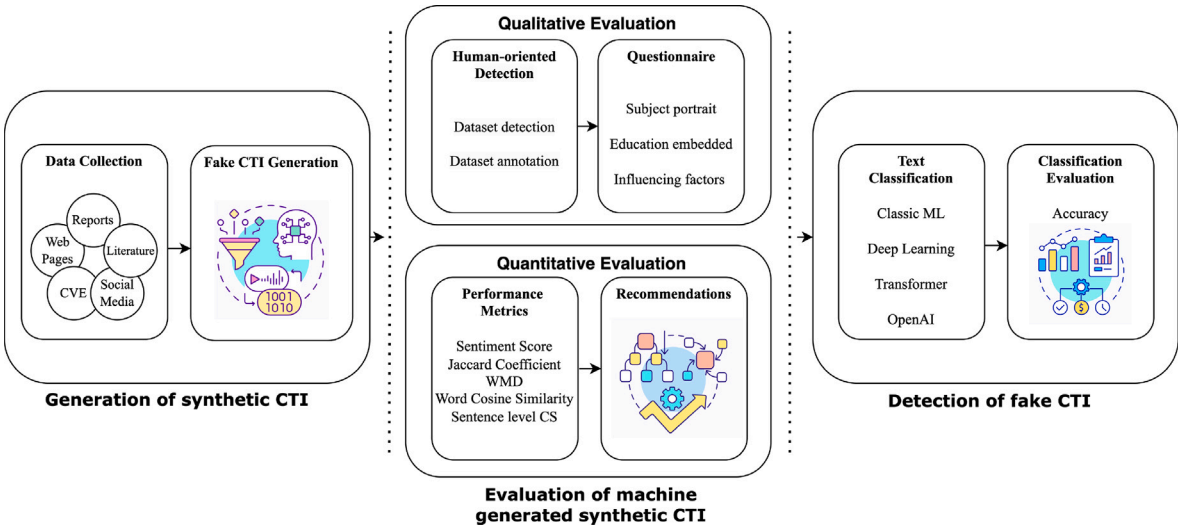


Fig. 1. Investigation Framework for LLM-Generated CTI. The framework consists of four major components: (1) Data Collection, where cybersecurity-related text from various sources (e.g., reports, social media, CVEs) is gathered; (2) Fake CTI Generation using LLMs trained on domain-specific data; (3) Evaluation, which includes both qualitative (human-oriented) and quantitative (statistical) validation of the generated CTI; and (4) Detection, where multiple machine learning and deep learning classifiers assess the credibility of CTI samples.

2.4.2. Automatic detection methods

The evolution of automated misinformation detection has progressed from traditional machine learning approaches to modern deep learning-based methods. Classical classifiers, such as logistic regression, Naive Bayes, and Support Vector Machines (SVM), have historically been used for misinformation classification [38,39]. However, neural misinformation models have introduced new challenges, requiring more robust, context-aware detection techniques. Studies have shown that the same models used to generate misinformation can often be the most effective at detecting their own output. Zellers et al. [40] demonstrated that models such as GPT-2 [41] and Grover [40] perform well in both misinformation generation and detection, suggesting that AI-generated text possesses distinct linguistic patterns detectable by similar architectures. Furthermore, transformer-based models such as Robustly Optimized BERT Pretraining Approach (RoBERTa) [42] have been widely recognized as state-of-the-art for detecting machine-generated text [38,41,43–45]. While these studies highlight advancements in general misinformation detection, they do not explicitly address CTI-specific misinformation. Existing research focuses primarily on news articles, social media content, and general AI-generated text, leaving a critical gap in applying misinformation detection to cybersecurity intelligence.

Unlike prior research, this study systematically evaluates multiple misinformation detection models within the CTI domain, comparing their effectiveness in distinguishing real and synthetic cybersecurity intelligence. The evaluated methods include traditional machine learning classifiers such as logistic regression and passive-aggressive models, neural network-based approaches such as Embeddings from Language Models (ELMo) [46] and GLTR [37], and transformer-based architectures such as RoBERTa and OpenAI's AI Text Classifier [47].

Existing studies in misinformation detection lack consensus on the most effective approach, with different models excelling in distinct scenarios. Some studies suggest statistical detection tools like GLTR are useful for human-assisted classification, while others argue that fine-tuned transformer-based models achieve superior generalization. However, no prior work has benchmarked these techniques for CTI-specific misinformation detection, making this research a foundational step in bridging this gap.

Compared to existing studies, this work differs in three fundamental ways. First, it focuses specifically on misinformation detection within Cyber Threat Intelligence rather than general fake news or AI-generated text classification. Second, it evaluates multiple misinformation detection models rather than focusing on a single architecture. Third, it addresses the unique challenges of CTI misinformation detection, such as domain-specific terminology and structured threat indicators. By identifying gaps in current misinformation detection research and adapting detection techniques to the cybersecurity domain, this study provides a foundation for developing AI-enhanced security tools to combat adversarial misinformation threats in cybersecurity.

3. Methodology

In this section, we provide a comprehensive description of the methodology we employed for generating, evaluating, and detecting fake CTI. Our approach encompasses several interconnected steps, as illustrated in Fig. 1.

Our methodology consists of three interconnected phases. First, we develop specialized CTI-focused language models to generate synthetic threat intelligence that closely mimics authentic CTI resources. Second, we implement a comprehensive evaluation framework that assesses the generated content across multiple dimensions, including semantic validity, contextual appropriateness, and conformity to established CTI patterns. Our evaluation confirms that the synthetic CTI matches authentic CTI in both linguistic patterns and structural characteristics, with human analysts unable to reliably distinguish between the two. Finally, we benchmark the detection of synthetic CTI using our groundtruth dataset across various detection methods. We maintain an iterative approach throughout, continuously refining our generation, evaluation, and detection methods based on empirical results and emerging insights.

3.1. Generating synthetic CTI

3.1.1. Data collection

Our analysis employed two complementary datasets: CASIE and UMBC CyberBlogDataset.

CASIE [24] provides long-form CTI content through 1000 expert-validated English cybersecurity news articles. Its semantic model captures diverse CTI elements including vulnerabilities and cyberattacks, organizing them into comprehensive knowledge graphs. This dataset offers detailed coverage of various cybersecurity incidents and their underlying vulnerabilities.

The UMBC CyberBlogDataset [48] supplies short-form CTI content from expert-annotated cybersecurity blogs. This dataset is structured in two formats: paragraph-level text files and sentence-level segments created using SpaCy sentencizer [49].

These datasets provide complementary perspectives - CASIE offering in-depth incident analysis through news articles, and UMBC delivering concise technical insights from cybersecurity blogs. This combination enables comprehensive analysis across different CTI content lengths and styles.

3.1.2. Synthetic CTI generation

Our data preprocessing strategy incorporated specialized techniques to optimize the model's comprehension and generation capabilities. Following Lee et al.'s [50] approach, we enhanced the CASIE dataset by adding the @@@ symbol after each initial sentence to demarcate generation prompts, and appended <end of text> tokens to signal sequence completion.

The preprocessed CASIE dataset was partitioned into an 80–20 split for training and testing, respectively. We employed GPT-2 (1.5B parameters) as our foundation model, leveraging its pre-training on WebText's 8 million web pages [51], which included baseline cybersecurity knowledge. Two separate fine-tuning processes were implemented: (1) Long-form CTI Generation: The model was fine-tuned on the CASIE dataset to generate comprehensive threat intelligence narratives; (2) short-form CTI Generation: A separate fine-tuning process using the UMBC CyberBlogDataset's Sentence category enabled the model to produce concise technical CTI content.

This domain-adaptive intermediate fine-tuning approach enhanced the model's capability to generate contextually relevant cybersecurity content across varying lengths and complexities. The dual fine-tuning strategy enabled our model to capture both detailed threat narratives and precise technical descriptions essential for comprehensive CTI generation.

To ensure that our dataset effectively represents the diversity of real-world CTI, we have categorized the collected and generated samples into seven primary cybersecurity themes. A brief description of each category, along with representative sample's topic examples, is provided below:

1. *Data Breaches and Leaks:* Incidents where organizations experience unauthorized access and exposure of sensitive data.
   Example: The Equifax breach, where personal data of 147 million individuals was compromised due to an unpatched vulnerability.
2. *Ransomware and Malware Attacks:* Cyber threats involving malicious software that encrypts files or disrupts systems.
   Example: The WannaCry ransomware attack, which infected over 200,000 computers worldwide and demanded Bitcoin payments.
3. *Phishing and Social Engineering Attacks:* Cybercriminals manipulating users into providing confidential information through deceptive means.
   Example: Tax refund phishing scams where attackers impersonate tax authorities to steal sensitive financial data.
4. *Critical Vulnerabilities and Exploits:* Identification and exploitation of software/hardware vulnerabilities.
   Example: Spectre and Meltdown CPU vulnerabilities that allowed attackers to extract sensitive data from affected systems.
5. *Cybercrime and Dark Web Activities:* Illegal transactions and cybercriminal activities in underground forums.
   Example: Stolen PlayStation accounts being sold on the dark web, leading to financial fraud.
6. *Government and Corporate Cybersecurity Issues:* Cybersecurity incidents impacting public sector entities or major corporations.
   Example: The City of Atlanta ransomware attack, which disrupted municipal services for weeks.
7. *Fake cybersecurity news, CTI Misinformation, and AI-Generated Threats:* The creation and dissemination of false cybersecurity information to mislead organizations.
   Example: Threat actors fabricating fake CTI reports to manipulate security responses and divert attention from real threats.

By incorporating these categories, our dataset offers a comprehensive representation of CTI challenges. Finally, we named our generated dataset GFCTI dataset.

3.2. Assessing the quality and authenticity of the generated CTI

3.2.1. Qualitative evaluation

Our assessment method, which was approved by the Human Research Ethics Committee of the University of New South Wales (HC220661), involved 125 participants completing a structured questionnaire as part of our crowdsourcing approach. This questionnaire was designed to investigate three key aspects of CTI authenticity:

- RQ1 (Assessment): How does the quality and authenticity of LLM-generated Cyber Threat Intelligence (CTI) compare to human-authored content in terms of technical accuracy, contextual relevance, and narrative coherence?
- RQ2 (Knowledge Impact): To what extent does an individual's IT expertise influence their ability to distinguish between synthetic and authentic CTI? Does technical proficiency correlate with higher detection accuracy?
- RQ3 (Decision Factors): What cognitive and contextual factors affect human judgment when identifying synthetic cybersecurity intelligence? How do elements such as writing style, technical depth, and contextual consistency influence detection decisions?

3.2.2. Quantitative evaluation

To systematically assess the quality and authenticity of LLM-generated CTI, we employ a set of quantitative metrics designed to capture lexical, semantic, and contextual properties of text. These metrics allow for an objective comparison between synthetic and real CTI, helping to determine the extent to which LLM-generated CTI mimics genuine cybersecurity intelligence. Below, we detail the performance metrics used to evaluate the generated text:

- *Sentiment Score:* We use TextBlob [52] to assign sentiment polarity to both synthetic and real CTI. This helps determine whether generated CTI maintains an appropriate emotional tone that aligns with real-world reports.
- *Jaccard Coefficient:* Measures the overlap between tokenized sets of words in real and synthetic CTI. This lexical similarity metric is chosen because keyword relevance is crucial in cybersecurity reports, where specific terms (e.g., "vulnerability", "exploit", "ransomware") indicate authenticity.
- *Word Mover's Distance (WMD):* Evaluates the semantic similarity between real and generated CTI using word embeddings [53]. Unlike lexical similarity metrics, WMD quantifies how closely the generated text maintains meaning by comparing it to existing cybersecurity intelligence. To better represent the similarity between word vectors, we utilize three pre-trained word2vec models from relevant domains as listed below:

  1. *GoogleNews-vectors-negative300* [54]: is a common embedding that contains 300 million commonly used word vectors trained by Google based on the large GoogleNews corpus. Each word vector has 300 dimensions.
  2. *Domain-word2vec* [24]: employs 100-dimensional randomly initialized vectors. It utilizes a vocabulary comprising 28,283 cybersecurity words.
  3. *Cyber-word2vec* [25]: is trained on a substantial corpus comprising approximately 1 million web pages related to cybersecurity, which includes a vocabulary of 6,417,554 words. Each word vector has 100 dimensions.

  The use of multiple embeddings ensures robustness in capturing both general linguistic and cybersecurity-specific semantic relationships.

- *Word Cosine Similarity:* In evaluating the quality of generated CTI, it is essential to measure how closely the generated text aligns with its original prompt (Topic). Word cosine similarity serves as a key semantic metric in our evaluation framework. Unlike Word Mover's Distance (WMD), it is less sensitive to variations in text length. It measures the cosine of the angle between two vectors. The smaller the angle, the larger the cosine value, which means the similarity between the two words is higher. Specifically, the closer the cosine value is to 1, the more similar the two texts are. To ensure a comprehensive evaluation, we employ two calculation methods, Scikit-learn [55] and SpaCy [56], as listed below.

    1. *Scikit-learn* [55]: We use Scikit-learn text representation model Term Frequency-Inverse Document Frequency (TF-IDF) to convert documents into vectors, and then calculate the cosine similarity respectively for the true and fake CTI content and topic.
    2. *SpaCy* [56]: Pauzi et al. [57] demonstrate that SpaCy exhibits outstanding performance in comparing text similarity, where captures semantic relationships beyond mere word overlap. For our analysis, we utilize the word embeddings and cosine function provided by the medium-sized SpaCy model.

- *Sentence-Level Cosine Similarity:* Captures structural similarity beyond the lexical level by comparing entire sentences instead of individual words. We utilize Sentence-BERT (SBERT) [58] for this purpose, as prior research has shown that transformer-based embeddings effectively represent sentence-level meaning [59].

To visualize how these metrics differentiate real and synthetic CTI, we generate density plots of their distributions. Furthermore, we employ a logistic regression model incorporating these metrics as features to quantify the alignment between generated CTI and real samples. The ranking of feature importance is determined using Variable Ranking (VR) and Maximal Information Coefficient (MIC), which highlight the most discriminative metrics. By integrating these diverse evaluation methods, we ensure a comprehensive assessment of LLM-generated CTI, providing insights into its authenticity, relevance, and potential risks in cybersecurity applications.

3.3. Identifying synthetic CTI

To establish comprehensive benchmarks for fake CTI detection, we evaluated multiple detection approaches: three traditional machine learning models as baseline classifiers, enhanced detection models based on GLTR [37] and ELMo [46] architectures, a RoBERTa transformer model fine-tuned on GPT-2 outputs, and OpenAI's AI text classifier [47]. This diverse set of models allows us to compare the effectiveness of both classical machine learning techniques and advanced deep learning approaches in identifying fake CTI.

For generating comprehensive benchmarks, we curated a balanced test set of 180 fake CTI samples from our dataset, each containing more than 1000 characters (equivalent to 164+ words). The samples were stratified across six word-count ranges (150–199, 200–249, 250–299, 300–349, 350–399, and 400–449 words), enabling us to benchmark both overall detection accuracy and analyze how text length influences model performance across different detection approaches.

3.3.1. Classic machine learning methods

For CTI sample detection, we employ three widely used classical machine learning algorithms as our baseline models. These algorithms are the Logistic Regression Classifier, Passive Aggressive Classifier, and Random Forest Classifier. The input for these models is the TF-IDF vector representation.

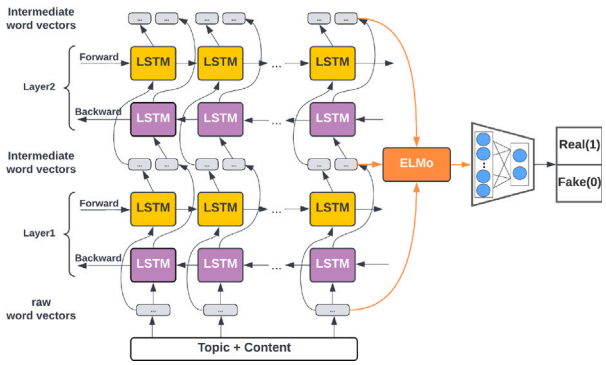


Fig. 2. The structure of our ELMo-based synthetic CTI detection method.

3.3.2. Classification based on ELMo

The primary drawback of global word embeddings, for example, TF-IDF and Word2Vec, is their failure to capture contextual meaning. They struggle with the issue of polysemy, where a word or phrase can have multiple possible meanings. For instance, they cannot distinguish whether 'Apple' refers to a fruit or a computer brand. In sentences where words have different meanings, they require distinct representations in the embedding space.

To address this limitation, contextual embedding methods such as BERT and Embeddings from Language Models (ELMo) [46] have been developed. These methods learn sequence-level semantics by considering the word order within a document. ELMo, for instance, computes embeddings based on the internal states of a two-layer bidirectional Language Model (LM). Unlike traditional word embeddings, ELMo generates multiple word embeddings for a single word in different contexts. It assigns each word a representation that is a function of the entire corpus of sentences, dynamically creating embeddings as needed. The embeddings are derived from all the internal layers of the bi-LSTM. The lower-level bi-LSTM layer captures syntactic information, while the higher-level bi-LSTM layer extracts semantic information. By concatenating the activations of all layers, ELMo can combine a diverse range of word representations, leading to better performance in downstream tasks.

Furthermore, ELMo operates at the character level rather than the word level. This allows it to utilize sub-word units to generate meaningful embeddings even for words not present in its vocabulary. In our approach, we employ ELMo to create input representations. For the representation of the CTI sample sequence $X_i, i = (1, 2, 3, \ldots, n)$, the formula is as follows:

$$X_k = \gamma \sum_{j=0}^{L} Sj h_{k,j}^{LM} \tag{1}$$

where $j = 0, 1, 2 \ldots L$, j indicates the jth layer, and $j = 0$ is the input layer. k means word position, and h means whole, that is, at each position k, each biLSTM layer (LM) will generate a representation of the word vector $h_{k,j}^{LM}$. Sj indicates the probability value after softmax, which can be understood as the weight of the output of each layer. The constant parameter γ is the scaling parameter used to scale the entire ELMo word vector. These two parameters are hyperparameters that need to be learned.

Finally, we feed the representation to a linear classifier to obtain the desired output. The structure of our model is depicted in Fig. 2.

3.3.3. Classification based on GLTR

Giant Language Model Test Room (GLTR) [37] was developed to identify machine-generated text by statistically analyzing and visualizing the given text. An example of GLTR's application of CTI is shown in Fig. 3. The key concept behind GLTR's detection of generated text is to use a similar model to the one originally used to generate the text, even

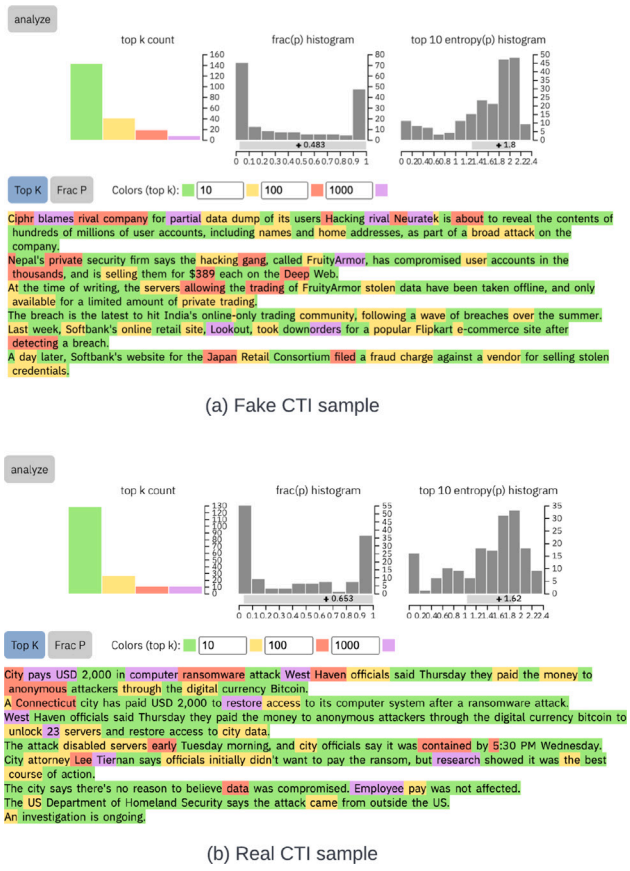


Fig. 3. GLTR-based analysis of Fake and Real CTI samples. The top section of each panel contains three statistical graphs: **Top-k count:** Indicates the frequency of words appearing in the most probable k-ranks predicted by the language model. **Frac(p) histogram:** Displays the fraction of words falling within different probability ranges. **Entropy histogram:** Represents the unpredictability of the word distribution in the sample. The lower part highlights the CTI text with top-k color-coded overlays, where: *Green (top 10):* Highly predictable tokens. *Yellow (top 100):* Moderately predictable. *Red (top 1000):* Less predictable. *Purple (outside top 1000):* Highly unlikely under normal language distribution. Comparing (a) Fake CTI and (b) Real CTI, we observe that the synthetic CTI samples closely resemble real samples in token distribution and color annotation, indicating that fake CTI generated by LLMs can be highly deceptive..

though the language model generates words based on the probability distribution it has learned from the training data. Gehrmann et al. [37] demonstrated that the annotation scheme provided by GLTR increased human detection of fake text from 54% to 72% without any training. By employing techniques, including maximum sampling, k-max sampling, beam search, kernel sampling, etc., it is possible to verify if words in a given text conform to a specific distribution. If multiple words in the text adhere to such a distribution, it suggests that the text is likely machine-generated.

Based on this principle, we transform GLTR, which utilizes a pre-trained GPT-2 model, into an automatic detection model. The specific method involves processing and training the input text using the pre-trained GPT2Tokenizer and GPT2LMHeadModel [60] classes. The GLTR model's probability formula is then utilized for calculation. The output includes four values: y_1, y_2, y_3, and y_4, representing the occurrence frequency ranges: less than 10%, 10% to 100%, 100% to 1000%, and more than 1000%, respectively. Finally, a linear classification is performed to classify the text as real or fake. The detailed structure is illustrated in Fig. 4.

3.3.4. Transformer based models

Robustly Optimized BERT Pretraining Approach (RoBERTa) [42], an improved method proposed by Facebook AI Research for training

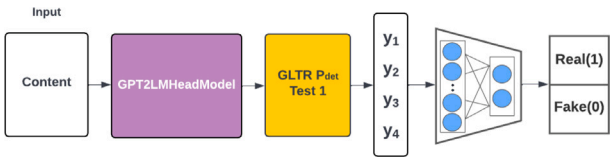


Fig. 4. The structure of our GLTR-based fake CTI detection method. The system consists of three key stages: **Text Embedding & Language Modeling:** Input content is processed through a pre-trained GPT-2 model. **GLTR Analysis:** The probability distribution of each token is computed, generating likelihood scores. **Classification Network:** Extracted statistical features (y_1, y_2, y_3, y_4) are passed through a deep neural network to classify CTI samples as real (1) or fake (0). The integration of GLTR-based linguistic analysis and neural network classification enhances the detection of fabricated CTI.

BERT models, has shown superior performance compared to the subsequent methods developed after BERT. In our study, we utilize the large RoBERTa model, which is fine-tuned using the output of the 1.5 billion parameters GPT-2 model [41]. This specific model is designed to detect text generated by GPT-2. We evaluate the input by testing all 2000 instances of the content in our dataset.

3.3.5. Web interface model and AI text classifier

The AI Text Classifier [47], developed by OpenAI, is a model fine-tuned on a pre-trained language model with the aim of distinguishing between human-written and AI-generated text. The training process involves utilizing text generated by 34 models from 5 different organizations, including OpenAI. The datasets used for training included the new Wikipedia dataset, the WebText dataset collected in 2019, and the training set from InstructGPT. Balanced batches, consisting of equal proportions of AI-generated and human-written text, are used during training. The classifier assigns labels of 'very unlikely,' 'unlikely,' 'unclear,' and 'likely' to detect whether a document is AI-generated. The classifier demonstrates promising results with an AUC score of 0.97 on the validation set, and 0.66 on the challenge sets. However, OpenAI notes that the classifier is sensitive to text length [47].

4. Experiments and results

Below, we present our experimental results through the full pipeline of investigating whether LLM-generated CTI can be detected. Our analysis covers three main phases: CTI sample generation, generated CTI validation, and performance benchmarking of different detection methods.

4.1. Results of LLM generated synthetic CTI

4.1.1. Samples of our generated CTI dataset

Table 1 highlights representative examples of long-form CTI entries from our GFCTI dataset. These samples were visualized using GLTR in Fig. 3 to aid in analyzing their patterns. The token distribution patterns indicated by top-K color annotations of LLM-generated CTI closely mirror those found in authentic CTI samples. In addition, Table 2 presents examples of shorter CTI entries produced by our generation models.

4.2. Results of qualitative evaluation of synthetic CTI

4.2.1. Qualitative evaluation design

We first design a survey to assess the quality and authenticity of our LLM-generated synthetic CTI (i.e., fake CTI). The survey consists of three distinct sections: (1) CTI sample identification, where participants determine whether the given sample is true or false, (2) CTI sample annotation, where participants manually mark suspicious words/phrases in the provided CTI samples; and (3) CTI and IT knowledge effectiveness questionnaire, which includes a series of questions for participants

Table 1
Long text samples of our generated dataset.

Topic	Cyber Threat Intelligence	Real/Fake
Ciphr blames rival company for partial data dump of its users	Ciphr, a company which offers encrypted communications for BlackBerry 10 and Samsung Knox smartphones, claims that a rival firm are behind a data dump of its customers' email addresses and their device's IMEI numbers. A website displaying the alleged leaked data claims that "all Ciphr emails/servers have been compromised." Two sources that use Ciphr on their phones told Motherboard the leak includes their information as well as the data of other users. Specifically, the website lists users' email addresses and IMEI numbers, data which law enforcement can leverage to expose a user. In a message provided to Motherboard from one of its sources, the privacy platform says the data dump was not the result of a data breach. Instead Ciphr blames a rival company for the incident: "Our rapid growth has caught the attention of competitors seeking to slow us down by way of slander, blocking and DDOS [distributed denial of service attacks].... We were shocked that any company in this industry would release information to the public under any circumstance." Ciphr's management explains in a blog post that a rogue reseller who was granted access to its sales systems gave the information to SkySecure, which makes custom Blackberry devices. The company goes on to note that most of the information included in the data dump was already expired. But it does say a few active users' email addresses and IMEI numbers were included in the leak.	Fake
City pays USD 2000 in computer ransomware attack	West Haven officials said Thursday they paid the money to anonymous attackers through the digital currency Bitcoin. A Connecticut city has paid USD 2000 to restore access to its computer system after a ransomware attack. West Haven officials said Thursday they paid the money to anonymous attackers through the digital currency bitcoin to unlock 23 servers and restore access to city data. The attack disabled servers early Tuesday morning, and city officials say it was contained by 5:30 PM Wednesday. City attorney Lee Tiernan says officials initially did not want to pay the ransom, but research showed it was the best course of action. The city says there is no reason to believe data was compromised. Employee pay was not affected. The US Department of Homeland Security says the attack came from outside the US. An investigation is ongoing.	Real

Table 2
Short text samples of our generated dataset.

Sentence	Real/Fake
FastDecode password protected webshellding Multi-Platform Support As we developed our tool to support any platform, we realized that not everything was possible in all Windows versions.	Fake (generated)
Another interesting component of the Shamoon variant is the use of the SysInternals utility PSEXEC, which we reported on in May of 2021.	Fake (Generated)

to answer. After completing the survey, participants will be given the correct answers for their annotations. The main objectives of the first and second sections provide insights to address RQ1 as outlined in Section 3.2. In the third section, the questionnaire incorporates participants' profiles, embedded education, and factors that might influence their annotation decisions. Some participants are invited to participate multiple times, and different samples are presented for detection and annotation. This aims to address RQ2 as described in Section 3.2. The results obtained will be further analyzed in conjunction with participants' job functions to determine the impact of IT knowledge efficacy on fake CTI detection. Additionally, the questionnaire explores various factors influencing human judgments of fake cybersecurity news, addressing RQ3. We conduct a random selection of 100 samples from the GFCTI dataset, comprising 50 real samples and 50 fake samples. It is ensured that both real and fake samples are derived from non-repetitive topics. We illustrate the details of three sections of the survey below:

- CTI sample identification: The first section of the questionnaire consists of two single-choice questions. Participants will be randomly assigned 2 of the 100 CTI samples in the sample pool and judge whether they are real or fake CTIs.
- CTI sample annotation: CTI samples that participants had selected as fake in the first section would appear in the second section of their questionnaire, where they were invited to highlight parts (words, phrases, or sentences) that felt suspicious.
- CTI and IT knowledge effectiveness questionnaire: As shown in Table 3, the questionnaire consists of 25 questions. Behind these questions are three functional sections. The first section is a subject portrait, which is divided into two parts: 1-A Subject background survey, which is questions 1, 2, and 5; 1-B confidence survey, which is Questions 12 and 13. The second section, Education Embedded, mainly arouses subjects' awareness of fake news (questions 3 and 4) and prompts them to identify methods (questions 6–11) for understanding the subjects' confidence after embedded education. The third section (i.e., influencing factors) consists of 12 questions, with the first 10 questions falling into three main directions of influence identified by prior research, namely information readability (questions 16, 17, and 19), credibility (questions 18 and 22) [61], and confidence bias (questions 14, 15, 20, 21 and 23) [62]. Additional two text-entry items (24 and 25) are included, inviting subjects to write down their perceived factors.

4.2.2. Subjects

The job functions of 125 participants are categorized into three groups: (1) IT-related job functions, (2) non-IT job functions, and (3) unknown job functions. The first two categories of participants were recruited through MTurk, where individuals were able to complete the questionnaire multiple times. Participants from other sources (i.e., direct distribution using Qualtrics) were only allowed to complete the questionnaire once. This was done to investigate whether embedded education could enhance people's detection abilities. In total, the survey is conducted 146 times, as outlined in Table 4.

4.2.3. Results

We present the survey results to answer our three research questions defined in Section 3.2.

Response to RQ1: Assessment

Table 5 presents the results of human-oriented classification, demonstrating the accuracy of distinguishing between real and fake CTI. Our analysis reveals that the overall average accuracy of individuals

Table 3
Details of questionnaire.

Subjects portrait	Background	1. Have you heard of "fake news" before?
		2. Have you heard of "fake cyber threat intelligence" or "fake cybersecurity news" before?
		5. Which languages do you speak the most fluently?
	Confidence	12. Do you think you can filter all the fake CTI news by yourself?
		13. How would you rate your cybersecurity knowledge?
Education embedded	Awareness	3. How often do you notice fake cybersecurity-related information on the internet per week?
		4. How often have you been fooled by fake news thinking it is real news?
	Method	6. I check the legitimacy of a website before accessing it.
		7. I check the metadata of the image.
		8. I search the internet for the claims made in the article/image/post.
		9. I check the credibility of the author by reading other articles from him/her.
		10. I cross-check the references in the article.
		11. I cross-check the website data on the fact-checking websites.
Influencing factors	Readability	16. When you read cybersecurity news with an easy-to-understand title and coherent contents, do you feel more confident that it is true?
		17. When you read cybersecurity information that is supported by solid arguments (i.e., extensive and detailed explanation), do you feel more confident that it is true?
		19. When cybersecurity information is presented objectively and factual, do you feel more confident that it is true?
	Credibility	18. Does the cybersecurity information providing detailed explanations with previous similar examples make you feel more confident that it is true?
		22. When you read a cybersecurity article that reflects multiple viewpoints, do you feel more confident that it is true??
	Confidence bias	14. If the cybersecurity information is published by someone you trust, do you feel more confident that it is true?
		15. If the cybersecurity information is published by someone you have been following for a long time, do you feel more confident that it is true?
		20. When you read cybersecurity news that is in a way consistent with your personal belief impression, do you feel more confident that it is true?
		21. Do you feel more confident that the cybersecurity news you are reading is true if you have read similar ones before?
		23. How confident are you that the cybersecurity news you read is true if it is new to you?
	Supplementary factors	24. Other than the above points, what makes you believe the cybersecurity-related information (including cybersecurity articles, news, forum posts, etc.) is trustworthy?
		25. Other than the above points, what makes you believe the cybersecurity-related information (including cybersecurity articles, news, forum posts, etc.) is fake?

Table 4
Participation times and job functions of subjects.

Times	IT	Not IT	Unknown	Total
1	20	39	39	98
2	17	10	0	27
3	14	7	0	21
Total	51	56	39	146

Table 5
Accuracy of human-oriented detection.

Job function	Correct	Incorrect	Accuracy
Not IT	47	51	47.96%
IT	49	25	66.22%
Unknown	32	26	55.17%
Average	43	34	56.45%

identifying the LLM-generated CTI is merely 56.45%. Notably, among the participants, those working in the IT field exhibit the highest accuracy rate at 66.22%, while non-IT subjects achieve a significantly lower accuracy of only 47.96%. These findings strongly indicate that our generated CTI samples pose a challenge for both IT professionals and the general audience when it comes to differentiating between genuine and LLM-generated CTI.

Response to RQ2: Knowledge Impact

We conducted a longitudinal study with two participant groups, administering questionnaires at three intervals to assess their CTI detection capabilities. This design allowed us to track how detection accuracy evolved as participants gained experience through repeated exposure. Fig. 5 illustrates the learning progression and shows how professional background influenced participants' ability to acquire and apply CTI authentication skills.

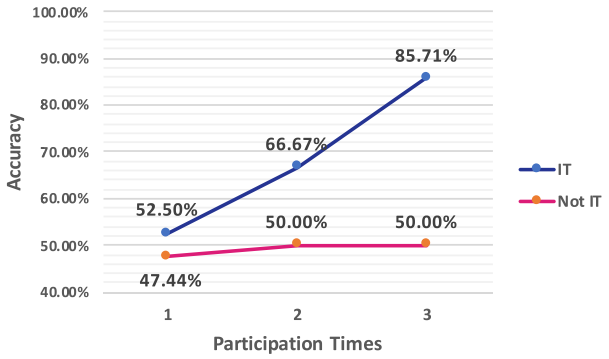


Fig. 5. Comparing three detection results of job function IT and Not IT.

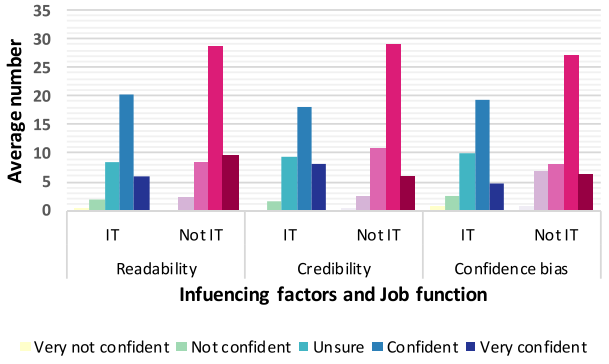


Fig. 6. Job function with influencing factors.

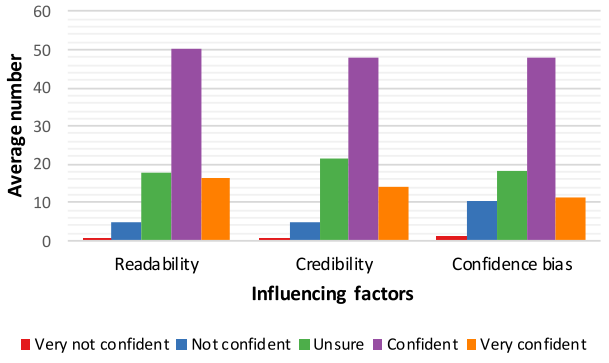


Fig. 7. Overall situation of influencing factors.

Fig. 5 reveals a stark contrast in learning trajectories between groups. While non-IT participants showed minimal improvement in detection accuracy across the three assessments, IT professionals demonstrated substantial progress, improving from 52.50% to 85.71% accuracy. This disparity suggests IT professionals' enhanced ability to acquire and apply cybersecurity knowledge. However, the initial similar performance between IT (52.50%) and non-IT participants (47.44%) indicates that without specific training, even IT professionals struggle to identify synthetic CTI. These findings emphasize the importance of developing accessible, non-technical training approaches for general audiences while highlighting the effectiveness of targeted training for technical professionals.

Response to RQ3: Decision Factors

We conducted an analysis to examine the factors influencing the human judgment of misinformation, including readability, credibility and confidence bias, and investigated their correlation with job functions (Fig. 6). Fig. 7 provides an overview of the subjects' choices concerning these three factors. In general, when information text is

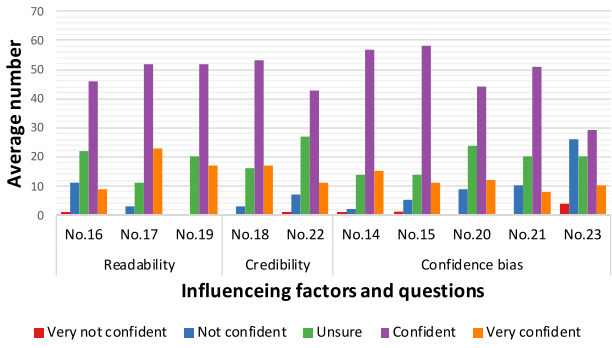


Fig. 8. Specific question of influencing factors.

highly readable, it tends to be perceived as true by a larger number of individuals. Examining specific questions, questions No. 15 and No. 14 emerged as the most influential factors (Fig. 8). These questions pertain to information published by individuals who are well-regarded or trusted over a long period, leading people to perceive it as genuine. This indicates that humans often exhibit a confidence bias when assessing the authenticity of information.

4.3. Results of quantitative evaluation of synthetic CTI

4.3.1. Results and visualization of designed performance metrics

We conducted a comprehensive analysis of the quantitative metrics (detailed in Section 3.2.2) from both authentic and LLM-generated CTI using density plots and feature importance analysis. The density plots visualize the distribution of various metrics including Sentiment Score, Jaccard Coefficient, Word Mover's Distance (three variants), Word Cosine Similarity (two variants), and sentence-level similarities. As shown in Fig. 9, our generated CTI closely mimics authentic samples across multiple dimensions, with sentence-level characteristics being nearly indistinguishable, as demonstrated in subfigure (h).

However, among all the metrics, word cosine similarity calculated using Scikit-learn proved to be the most effective in distinguishing between real and fake CTIs. The overall distribution of fake CTIs (red) is skewed to the left, indicating that LLM-generated CTIs exhibit lower lexical overlap with their original prompts compared to real CTIs. Since Scikit-learn's TF-IDF-based cosine similarity primarily captures word-level co-occurrence patterns, this suggests that LLM-generated CTIs tend to rephrase or substitute words more frequently, rather than strictly adhering to the exact vocabulary used in the prompts. Consequently, when using the SpaCy word embedding-based word cosine similarity method, which captures deeper semantic relationships, the distinction between real and fake CTIs becomes less apparent. This is because, while the generated content differs significantly from the original prompt in wording, it remains thematically consistent. Such characteristics may explain why even cybersecurity experts, despite their domain knowledge, struggle to differentiate between real and synthetic CTIs. These findings further highlight the necessity of employing multi-faceted evaluation techniques to enhance detection accuracy.

4.3.2. Recommendations of performance metrics of evaluating LLM-generated text

To further assess the discriminative power of our quantitative metrics (Section 3.2.2), we implemented a dual-ranking methodology. The first approach employed Variable Ranking (VR), combining F-test statistics with K-best feature selection to identify statistically significant features. The second utilized the Maximal Information Coefficient (MIC), a sophisticated measure from the MINE statistics family that captures both linear and nonlinear relationships between variables, offering a more comprehensive understanding of feature importance.

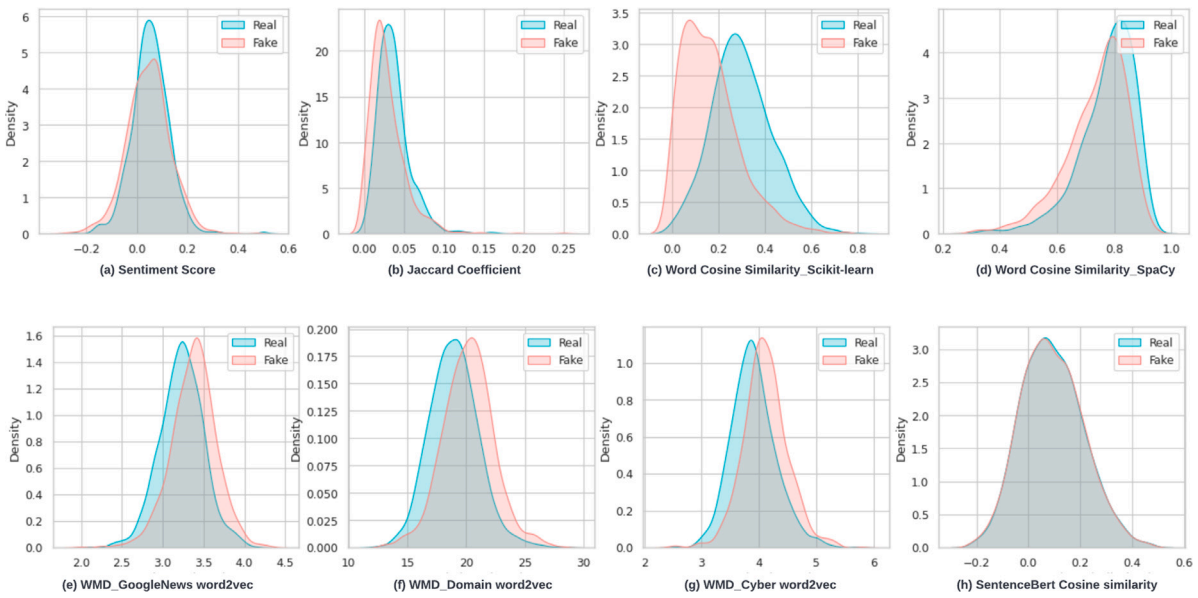


Fig. 9. Density plot of metrics. On the top, show four metrics (a) Sentiment Score, (b) Jaccard Coefficient, (c) Word Cosine Similarity Scikit learn, and (d) Word Cosine Similarity Spacy learn. On the bottom, show others four metrics, (e) WMD GoogleNews word2vec, (f) WMD Domain word2vec, (g) WMD Cyber word2vec, and (h) SentenceBert Cosine similarity.

Table 6
Performance metrics recommendation.

Metrics		Score			Accuracy		
		VR	Rk	MIC	Rk	Value	Rk
Sentiment		4.83	7	0.1129	7	50.00%	7
Jaccard		66.61	6	0.1502	5	60.50%	5
WCS	SKlearn	578.13	1	0.2951	1	71.25%	1
	SpaCy	68.98	5	0.1204	6	53.00%	6
WMD	Google	142.44	3	0.1516	4	65.00%	2
	Domain	165.41	2	0.1597	3	62.50%	3
	Cyber	137.60	4	0.1600	2	60.75%	4
SCS	SBERT	0.02	8	0.0885	8	47.25%	8

Table 7
Performance on detection models.

Type	Models	Accuracy	Precision	Recall	F1-score
Machine Learning	Logistic Regression	0.5925	0.5813	0.6020	0.5915
	Passive Aggressive	0.5775	0.5616	0.5876	0.5743
	Random Forest	0.6800	0.7241	0.6712	0.6967
Deep Learning	Based-ELMo	0.7225	0.7016	0.7128	0.7071
	Based-GLTR	0.5725	0.5359	0.6022	0.5671
Transformer	RoBERTa	0.9370	0.8780	0.9955	0.9330
Web Interface	AI Text Classifier	0.1100	N/A	N/A	N/A

Furthermore, we integrated these metrics into a logistic regression framework for binary classification between authentic and LLM-generated CTI. Table 6 reveals a clear hierarchy of feature importance across both ranking methods. Word-level similarity metrics, particularly those computed using Scikit-learn (SKlearn), demonstrated superior discriminative power in both VR and MIC rankings. In contrast, transformer-based sentence similarity measures (SBERT) showed limited effectiveness in distinguishing between real and synthetic CTI. This suggests that lexical-level features may be more reliable indicators of synthetic content than semantic-level representations.

4.4. Performance on fake CTI detection approaches

4.4.1. Performance on classic machine learning models

Our evaluation of traditional machine learning approaches included Logistic Regression, Passive Aggressive, and Random Forest classifiers (Table 7). While all three methods performed marginally above random chance, Random Forest demonstrated the strongest performance with 68% accuracy, followed by Logistic Regression at 59.25% and Passive Aggressive at 57.75%. These results suggest that while classical machine learning techniques can detect some patterns in LLM-generated CTI, their effectiveness is limited.

4.4.2. Performance on deep learning and transformer approaches

We conducted a performance comparison of four detectors, which included two automatic detectors based on the deep learning models ELMo and GLTR, the transformer-based RoBERTa, and the latest AI-generated text detector developed by OpenAI. For the ELMo and GLTR models, we trained them using 80% of the available data and tested them using the remaining 20%. As for RoBERTa GPT-2, it was fine-tuned on GPT-2 generated text and directly tested on the dataset. Additionally, we manually performed the detection using the OpenAI web interface. The outcomes of the four models are displayed in Table 7.

RoBERTa, fine-tuned based on GPT-2, exhibited exceptional performance by achieving an accuracy of 93.65% even without prior exposure to the LLM-generated CTI dataset. This outcome suggests that detecting distinct features left by generative model architectures in fake texts holds promise for effective detection with potential generalizability. However, a potential challenge lies in real-world scenarios where the generative model of the fake CTI remains unknown. Furthermore, the based-ELMo model achieved an accuracy of 72.25%, ranking second in our comprehensive study. This finding highlights the crucial role played by the text vectorization method in text detection.

On the other hand, the performance of the based-GLTR model did not significantly outperform random values. This result indicates that modern generative models are increasingly capable of mimicking human-like vocabulary choices, thereby narrowing the gap in diversity and breadth. As for the AI text classifier, it achieved an accuracy rate of only 1.1%, significantly lower than the 26% accuracy claimed by OpenAI for an unknown challenge set. Given its heavy reliance on text length, we investigated the impact of text length on its performance, and the results are depicted in Fig. 10.

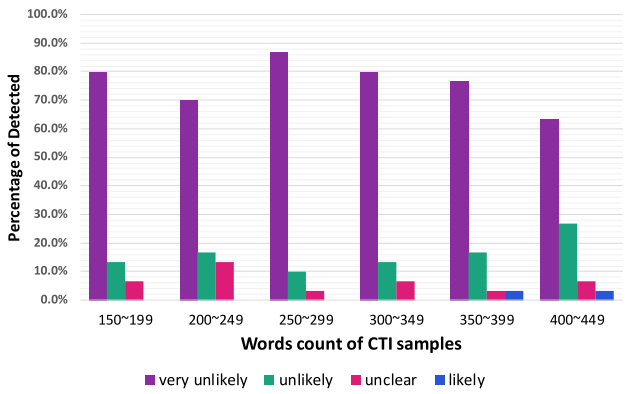


Fig. 10. Comparing the performance of classifiers for different text lengths. The results show that how likely is the CTI sample that was generated by AI.

Analysis reveals a correlation between text length and detection confidence. Longer texts are more frequently identified as AI-generated, with the classifier showing the increased probability of detecting synthetic content. However, the classifier's confidence in identifying human-authored text decreases with length, demonstrated by a declining proportion of 'very unlikely' AI-generated classifications and a corresponding rise in 'unlikely' determinations. This pattern demonstrates that text length is a significant factor influencing the detectors' performance.

5. Discussion and conclusion

5.1. Real-world integration and practical applications

Our proposed framework offers significant practical value for real-world cybersecurity applications by addressing the growing threat of fake CTI. By leveraging the generated dataset and dual validation framework, organizations can build automated systems for detecting fabricated CTI and integrate them into existing workflows. Specifically, the validated dataset can be used to train domain-specific machine learning models, which can be incorporated into Security Information and Event Management (SIEM) platforms or Threat Intelligence Platforms (TIPs) to automatically flag suspicious or potentially fake CTI reports. This enables cybersecurity teams to focus on high-confidence intelligence, reducing the risk of acting on malicious or misleading information.

Furthermore, organizations can adopt the dual validation framework to establish in-house pipelines for CTI assessment. Human validation, performed by cybersecurity professionals, can ensure the accuracy and relevance of critical intelligence reports, while statistical metrics provide scalable, automated screening to handle large volumes of CTI. By combining human expertise with statistical evaluation, organizations can strike a balance between efficiency and accuracy, improving their ability to detect and mitigate the risks associated with fake CTI.

The framework's flexibility also allows for the development of tools tailored to specific needs, such as lightweight detection systems utilizing statistical metrics to assess the quality of CTI. These tools can serve as early warning systems, flagging reports that deviate significantly from trusted sources. Additionally, the insights from the framework can be incorporated into training programs for cybersecurity analysts, helping them identify linguistic patterns and inconsistencies commonly associated with fake CTI.

By providing a modular and adaptable solution, the proposed framework can be customized to meet the needs of various organizations, whether they prioritize automation, human oversight, or a hybrid approach. Integrating this framework into existing cybersecurity workflows enhances the ability of practitioners to detect and mitigate fake CTI, ultimately improving organizational resilience against evolving cyber threats.

While improving detection techniques for fake CTI is critical, our findings indicate that detection alone is insufficient to fully mitigate the risks posed by misinformation in cybersecurity workflows. The fact that both humans and machine learning models struggle to reliably distinguish synthetic CTI from real intelligence highlights a fundamental vulnerability in current CTI processing pipelines. To strengthen the integrity of CTI beyond detection, provenance tracking and verification mechanisms should be integrated into cybersecurity workflows. Organizations can implement cryptographic verification techniques (e.g., digital signatures or blockchain) to authenticate intelligence sources and prevent tampering. Additionally, cross-referencing intelligence from multiple independent sources can enhance the credibility of CTI reports and reduce the impact of misinformation. A hybrid validation approach, such as the one proposed in this study, where statistical verification is complemented by expert human review, provides a scalable method for improving CTI reliability.

Ultimately, our results emphasize the importance of multi-layered defenses against fake CTI. Rather than relying solely on detection models, organizations should adopt a comprehensive approach that incorporates verification, provenance tracking, and human oversight. By combining these strategies, cybersecurity professionals can mitigate the risks associated with AI-generated misinformation and enhance the trustworthiness of threat intelligence in real-world applications.

5.2. Conclusion

The validation of Cyber Threat Intelligence (CTI) quality is still in its early stages, and there is a critical need for high-quality groundtruth datasets to develop effective fake CTI detection. This is crucial for further integrating CTI into intrusion prevention systems and security information and event management systems, where accurate detection of fake CTI is paramount. To address this need, this study presents a validated dataset specifically designed to identify fake CTI information, enabling the accurate identification and filtering of misinformation in the realm of cybersecurity. To evaluate the quality and authenticity of machine-generated text, particularly machine-generated CTI, we introduced innovative approaches that incorporate quality indication metrics. By leveraging the validated dataset, which encompasses both real and fake CTI samples, we conducted comprehensive evaluations of various state-of-the-art advanced detection models. These findings serve as a valuable reference for automating the detection of fake CTI, providing guidance for the development of robust detection systems. Additionally, we explored human-oriented detection methods and examined the real-world implications of incorporating embedded education. This research offers insightful guidance for implementing measures to combat fake CTI within organizations, contributing to enhancing overall cybersecurity practices and strategies.

CRediT authorship contribution statement

He Huang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Nan Sun:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Massimiliano Tani:** Writing – review & editing, Validation, Supervision. **Yu Zhang:** Writing – review & editing, Validation, Supervision. **Jiaojiao Jiang:** Writing – review & editing, Supervision. **Sanjay Jha:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This paper is supported by the UNSW AI Seed funding.

Data availability

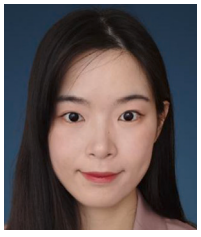
We will make both the data and code available on GitHub in a public repository.

References

[1] M.S. Abu, S.R. Selamat, A. Ariffin, R. Yusof, Cyber threat intelligence–issue and challenges, Indones. J. Electr. Eng. Comput. Sci. 10 (1) (2018) 371–379.
[2] N. Sun, M. Ding, J. Jiang, W. Xu, X. Mo, Y. Tai, J. Zhang, Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives, IEEE Commun. Surv. Tutor. (2023).
[3] H.M. Alzoubi, T.M. Ghazal, M.K. Hasan, A. Alketbi, R. Kamran, N.A. Al-Dmour, S. Islam, Cyber security threats on digital banking, in: 2022 1st International Conference on AI in Cybersecurity, ICAIC, IEEE, 2022, pp. 1–4.
[4] Y. Gao, X. Li, H. Peng, B. Fang, S.Y. Philip, Hincti: A cyber threat intelligence modeling and identification system based on heterogeneous information network, IEEE Trans. Knowl. Data Eng. 34 (2) (2020) 708–722.
[5] O. Kayode-Ajala, Applications of Cyber Threat Intelligence (CTI) in financial institutions and challenges in its adoption, Appl. Res. Artif. Intell. Cloud Comput. 6 (8) (2023) 1–21.
[6] Z. Song, Y. Tian, J. Zhang, Y. Hao, Generating fake cyber threat intelligence using the gpt-neo model, in: 2023 8th International Conference on Intelligent Computing and Signal Processing, ICSP, IEEE, 2023, pp. 920–924.
[7] Z. Li, X. Yu, Y. Zhao, A web semantic mining method for fake cybersecurity threat intelligence in open source communities, Int. J. Semant. Web Inf. Systems (IJSWIS) 20 (1) (2024) 1–22.
[8] N. Sun, J. Zhang, S. Gao, L.Y. Zhang, S. Camtepe, Y. Xiang, Cyber information retrieval through pragmatics understanding and visualization, IEEE Trans. Dependable Secur. Comput. 20 (2) (2022) 1186–1199.
[9] P. Ranade, A. Piplai, S. Mittal, A. Joshi, T. Finin, Generating fake cyber threat intelligence using transformer-based models, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–9.
[10] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, J. Artificial Intelligence Res. 61 (2018) 65–170.
[11] G. Cascavilla, D.A. Tamburri, W.-J. Van Den Heuvel, Cybercrime threat intelligence: A systematic multi-vocal literature review, Comput. Secur. 105 (2021) 102258.
[12] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, R. Beyah, Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 755–766.
[13] P. Gao, F. Shao, X. Liu, X. Xiao, Z. Qin, F. Xu, P. Mittal, S.R. Kulkarni, D. Song, Enabling efficient cyber threat hunting with cyber threat intelligence, in: 2021 IEEE 37th International Conference on Data Engineering, ICDE, IEEE, 2021, pp. 193–204.
[14] S. Barnum, Standardizing cyber threat intelligence information with the structured threat information expression (stix), Mitre Corp. 11 (2012) 1–22.
[15] 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: Proceedings of the 33rd Annual Computer Security Applications Conference, 2017, pp. 103–115.
[16] N. Sun, J. Zhang, S. Gao, L.Y. Zhang, S. Camtepe, Y. Xiang, Data analytics of crowdsourced resources for cybersecurity intelligence, in: Network and System Security: 14th International Conference, NSS 2020, Melbourne, VIC, Australia, November 25–27, 2020, Proceedings 14, Springer, 2020, pp. 3–21.
[17] V. Orbinato, M. Barbaraci, R. Natella, D. Cotroneo, Automatic mapping of unstructured cyber threat intelligence: An experimental study, 2022, arXiv preprint arXiv:2208.12144.
[18] X. Bouwman, V. Le Pochat, P. Foremski, T. Van Goethem, C.H. Gañán, G.C. Moura, S. Tajalizadehkhoob, W. Joosen, M. Van Eeten, Helping hands: Measuring the impact of a large threat intelligence sharing community, in: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1149–1165.
[19] T.D. Wagner, K. Mahbub, E. Palomar, A.E. Abdallah, Cyber threat intelligence sharing: Survey and research directions, Comput. Secur. 87 (2019) 101589.
[20] W. Tounsi, H. Rais, A survey on technical threat intelligence in the age of sophisticated cyber attacks, Comput. Secur. 72 (2018) 212–233.
[21] M. Sarhan, S. Layeghy, N. Moustafa, M. Portmann, Cyber threat intelligence sharing scheme based on federated learning for network intrusion detection, J. Netw. Syst. Manage. 31 (1) (2023) 3.
[22] M. Allegretta, G. Siracusano, R. Gonzalez, M. Gramaglia, Are crowd-sourced CTI datasets ready for supporting anti-cybercrime intelligence? Comput. Netw. 234 (2023) 109920.
[23] G. Sakellariou, P. Fouliras, I. Mavridis, A methodology for developing & assessing CTI quality metrics, IEEE Access 12 (2024) 6225–6238.
[24] T. Satyapanich, F. Ferraro, T. Finin, Casie: Extracting cybersecurity event information from text, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 8749–8757.
[25] A. Padia, A. Roy, T.W. Satyapanich, F. Ferraro, S. Pan, Y. Park, A. Joshi, T. Finin, UMBC at SemEval-2018 task 8: Understanding text about malware, UMBC Comput. Sci. Electr. Eng. Dep. (2018).
[26] 2019 cyber-research, APT malware dataset, 2025, URL https://github.com/cyber-research/APTMalware?tab=readme-ov-file.
[27] W. Peng, J. Ding, W. Wang, L. Cui, W. Cai, Z. Hao, X. Yun, CTISum: A new benchmark dataset for cyber threat intelligence summarization, 2024, arXiv preprint arXiv:2408.06576.
[28] D. Kim, H.K. Kim, Automated dataset generation system for collaborative research of cyber threat analysis, Secur. Commun. Netw. 2019 (1) (2019) 6268476.
[29] J. Li, T. Tang, W.X. Zhao, J.-R. Wen, Pretrained language models for text generation: A survey, 2021, arXiv preprint arXiv:2105.10311.
[30] X. Shi, H. Huang, P. Jian, Y.-K. Tang, Improving neural machine translation with sentence alignment learning, Neurocomputing 420 (2021) 15–26.
[31] A. Alomari, N. Idris, A.Q.M. Sabri, I. Alsmadi, Deep reinforcement and transfer learning for abstractive text summarization: A review, Comput. Speech Lang. 71 (2022) 101276.
[32] W. He, Y. Dai, Y. Zheng, Y. Wu, Z. Cao, D. Liu, P. Jiang, M. Yang, F. Huang, L. Si, et al., Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 10749–10757.
[33] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N.A. Smith, Y. Choi, Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 7250–7274.
[34] Y. Qu, P. Liu, W. Song, L. Liu, M. Cheng, A text generation and prediction system: Pre-training on new corpora using bert and gpt-2, in: 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication, ICEIEC, IEEE, 2020, pp. 323–326.
[35] L. Dugan, D. Ippolito, A. Kirubarajan, C. Callison-Burch, Roft: A tool for evaluating human detection of machine-generated text, 2020, arXiv preprint arXiv:2010.03070.
[36] V.L. Rubin, On deception and deception detection: Content analysis of computer-mediated stated beliefs, Proc. Am. Soc. Inf. Sci. Technol. 47 (1) (2010) 1–10.
[37] S. Gehrmann, H. Strobelt, A.M. Rush, Gltr: Statistical detection and visualization of generated text, 2019, arXiv preprint arXiv:1906.04043.
[38] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J.W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, 2019, arXiv preprint arXiv:1908.09203.
[39] J. Bogaert, M.-C. de Marneffe, A. Descampe, F.-X. Standaert, Automatic and manual detection of generated news: Case study, limitations and challenges, in: Proceedings of the 1st International Workshop on Multimedia AI Against Disinformation, 2022, pp. 18–26.
[40] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, Adv. Neural Inf. Process. Syst. 32 (2019).
[41] G. Jawahar, M. Abdul-Mageed, L.V. Lakshmanan, Automatic detection of machine generated text: A critical survey, 2020, arXiv preprint arXiv:2011.01314.
[42] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[43] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: Conf. on Empirical Methods in Natural Language Processing, EMNLP, 2020.
[44] D.I. Adelani, H. Mai, F. Fang, H.H. Nguyen, J. Yamagishi, I. Echizen, Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection, in: International Conference on Advanced Information Networking and Applications, Springer, 2020, pp. 1341–1354.
[45] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, R. Srinivasan, Cross-domain detection of GPT-2-generated technical text, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 1213–1233.
[46] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018, http://dx.doi.org/10.48550/ARXIV.1802.05365, URL https://arxiv.org/abs/1802.05365.
[47] OpenAI, AI text classifier, 2023, URL https://beta.openai.com/ai-text-classifier.
[48] C. Hanks, M. Maiden, P. Ranade, T. Finin, A. Joshi, et al., Recognizing and extracting cybersecurity entities from text, in: Workshop on Machine Learning for Cybersecurity, International Conference on Machine Learning, 2022.
[49] SpaCy, SpaCy sentencizer, 2023, URL https://spacy.io/api/sentencizer.
[50] J.-S. Lee, J. Hsiang, Patent claim generation by fine-tuning openai GPT-2, World Pat. Inf. 62 (2020) 101983.
[51] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

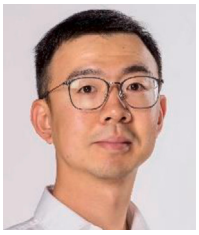
[52] W. Aljedaani, F. Rustam, M.W. Mkaouer, A. Ghallab, V. Rupapara, P.B. Washington, E. Lee, I. Ashraf, Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry, Knowl.-Based Syst. 255 (2022) 109780.
[53] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: International Conference on Machine Learning, PMLR, 2015, pp. 957–966.
[54] A. Giachanou, G. Zhang, P. Rosso, Multimodal fake news detection with textual, visual and semantic information, in: International Conference on Text, Speech, and Dialogue, Springer, 2020, pp. 30–38.
[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[56] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, To Appear. 7 (1) (2017) 411–420.
[57] Z. Pauzi, A. Capiluppi, Text similarity between concepts extracted from source code and documentation, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2020, pp. 124–135.
[58] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019, arXiv preprint arXiv:1908.10084.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[60] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[61] M. Maasberg, E. Ayaburi, C. Liu, Y. Au, Exploring the propagation of fake cyber news: An experimental approach, 2018.
[62] S. Suntwal, S. Brown, M. Patton, How does information spread? An exploratory study of true and fake news, 2020.

He Huang is a Ph.D. candidate in Computer Science at UNSW Canberra, specializing in artificial intelligence (AI). Previously, she served as a Research Assistant at UNSW Canberra and Deakin University, gaining over three years of experience in deepfake detection, disinformation detection, cybersecurity, and data analysis. She holds a Master's degree from Deakin University.

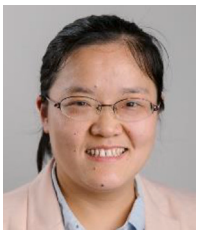


Dr Nan Sun received her Ph.D. degree in Information Technology from Deakin University. She is currently a lecturer with the School of Engineering and Information Technology at the University of New South Wales, Canberra. Before joining UNSW, she was a Research Fellow in the Centre for Cyber Security Research and Innovation (CSRI) at Deakin University and worked on the project - Development of Australian Cyber Criteria Assessment (DACCA). Her research focuses on cyber security, including data-driven cybersecurity incidents prediction, visualization and discovery through data analytics and machine learning techniques. Dr Sun is conducting interdisciplinary research between cybersecurity and artificial intelligence (AI), applying AI for cybersecurity. She is also passionate about designing systems to help with users' cybersecurity awareness education and cyber information retrieval.

Professor Massimiliano Tani Bertuol is an Academic in the School of Business at UNSW, Canberra. He is an economist by training and my research is applied. His interest focuses on human capital at large: how to foster it, its efficient international transfer through temporary and permanent migration, and its effects on productivity, innovation, and economic growth at a firm or national level. His education includes a Ph.D. in Economics from the Australian National University (Canberra, Australia), a M.Sc. Econ from the LSE and Laurea from Bocconi University (Milan, Italy).



Dr Zhang is a lecturer of data science at the School of Business, UNSW Canberra. His research interests contain text mining and analysis, knowledge and information management, social computing, and bibliometric analysis. He has applied his research to interdisciplinary areas and focused on publishing in prestige journals and conference, including Information Processing & Management, Journal of Informetrics, China Economic Review, Studies in Higher Education, AAAI, CIKM, and PAKDD, etc.



Dr Jiaojiao Jiang is currently a senior lecturer at the School of Computer Science and Engineering at the University of New South Wales. She holds a Ph.D. degree from Deakin University, Melbourne, Australia. She has published over 45 articles in high quality journals and conferences and received 1100 citations.

Her current research focuses on AI for Cybersecurity. In particular, she is interested in research at the detection of misinformation and modeling the propagation of misinformation on online social networks.

Jiaojiao has served as PC members of a number of conferences: ACM MM, CIKM, ECAI, etc. She has also been a regular reviewer for top ranked journals, including IEEE TIFS, IEEE TNSE, IEEE TDSC, etc.

Sanjay K. Jha is a full Professor at the School of Computer Science and Engineering since 2006. He is also the Director of Research and Innovation at the School of Computer Science and Engineering. He served as the Interim Director, Research Director and Chief Scientist of the UNSW Institute for Cybersecurity (IFCYBER). He holds a Ph.D. degree from the University of Technology, Sydney, Australia. Sanjay has published over 300 articles in high-quality journals and conferences. He is the principal author of the book Engineering Internet QoS and a co-editor of the book Wireless Sensor Networks: A Systems Perspective. He has been very active in attracting ARC Discovery and linkage grants, CRC and other industries. He leads UNSW's participation in the Cooperative Research Centre for Cyber Security (CSCRC).