Searching for Diamonds: Cross-Domain Opportunities in Cyber Threat Intelligence

Article *in* IEEE Access · January 2025

DOI: 10.1109/ACCESS.2025.3627126

5 authors, including:

Sidnei Barbieri

3 PUBLICATIONS **139** CITATIONS

SEE PROFILE

Flavio Luiz dos Santos de Souza

Federal Institute of São Paulo

4 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Marcio Andrey Teixeira

Federal Institute of São Paulo

40 PUBLICATIONS **993** CITATIONS

SEE PROFILE

Lourenco Alves Pereira Junior

The Brazilian Aeronautics Institute of Technology

65 PUBLICATIONS **392** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Searching for Diamonds: Cross-Domain Opportunities in Cyber Threat Intelligence

SIDNEI BARBIERI[1], FLAVIO LUIZ DOS SANTOS DE SOUZA[1,2], MARCIO ANDREY TEIXEIRA[2], CESAR AUGUSTO CAVALHEIRO MARCONDES[1], LOURENÇO ALVES PEREIRA JÚNIOR[1]

[1]Aeronautics Institute of Technology (ITA), São José dos Campos, SP, Brazil
[2]Federal Institute of São Paulo (IFSP), Catanduva, SP, Brazil

Corresponding author: Sidnei Barbieri (e-mail: sidneisb@ita.br).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Finance Code 001, and by the ITA's Programa de Pós-graduação em Aplicações Operacionais (ITA/PPGAO). The authors are also supported in part by the grant #2020/09850-0 from the São Paulo Research Foundation (FAPESP).

ABSTRACT Natural Language Processing (NLP) and Large Language Models (LLMs) are increasingly used in cybersecurity to enable automated, auditable, and intelligent systems. However, this convergence remains conceptually and methodologically underdeveloped in Cyber Threat Intelligence (CTI), a domain centered on processing large volumes of information. This paper presents a cross-domain review that synthesizes the state-of-the-art literature and outlines strategic opportunities, risks, and open challenges related to the application of LLMs in CTI. Guided by the conceptual lens of "searching for diamonds," the study identifies cross-cutting aspects—such as explainability, semantic reasoning, federated collaboration, model security, and machine unlearning—as essential to building normatively aligned autonomous CTI agents capable of efficiently processing the high-volume, dynamic data characteristic of CTI environments. It also highlights gaps in validation, trust, and regulatory compliance, emphasizing the need for co-evolutionary defenses and formal verifiability mechanisms. Grounded in multidisciplinary insights and aligned with international frameworks such as the Tallinn Manual and NIST/ISO standards, this work provides a strategic roadmap for future research and deployment of LLM-driven CTI architectures.

INDEX TERMS Cyber Threat Intelligence, Large Language Models, Natural Language Processing, Semantic Reasoning, Security Automation, Cybersecurity

I. INTRODUCTION

In recent decades, cybersecurity has evolved from a specialized technical field into a multidisciplinary domain with strategic impact on global security and governance [1]. Once restricted to experts, it now influences public discourse and strategic decisions by governments, industries, and civil society. This shift is driven primarily by the increasing digital interconnection of our lives, amplified by the widespread use of smart devices, cloud computing, and pervasive communication technologies, which have made the cyber environment a natural part of our lives. Our exposure to cyber threats increases as we increasingly rely on digital systems.

The public reaction to the blackout that affected Portugal and Spain in April 2025 exemplifies how cybersecurity concerns have expanded beyond the technical sphere and become issues of social relevance. Although Redes Energéticas Nacionais (REN), Portugal's national electricity grid operator, has publicly ruled out a cyberattack[1], the limited technical transparency of the investigation, combined with a growing global pattern of cyberattacks on critical infrastructure, has fueled public uncertainty and speculation. This response reflects an ever-increasing societal awareness that cyber threats are not mere technical anomalies but rather vulnerabilities that disrupt daily life and national security. The case highlights the complex interplay between cybersecurity, public trust, policymaking, and collective risk perception. This episode underscores the significance of reliable detection and attribution capabilities, particularly in systems characterized by low observability, outdated technologies, and proprietary protocols that impede independent audits.

In this rapidly evolving research landscape, understanding the current state of the art and recognizing new trends in cybersecurity results in a challenging and uncertain environment. Traditional methods that emphasize analysis and

[1]Reuters: *Portugal's REN says no blackout caused by cyberattack*

systematic organization of research are important for mapping scientific progress and identifying gaps. However, these methods are increasingly overwhelmed by the large volume of publications, varying terminologies, and the rapid pace of new knowledge production. Researchers face a challenge in synthesizing their work, identifying connections between unrelated subfields, and overcoming cognitive biases, such as confirmation bias, which can limit their exploration. Conventional research methods may unintentionally overlook emerging subdomains or subtle advances in rapidly evolving fields, especially when constrained by time, resources, or the need for upfront screening processes [2]. As a result, there is a high risk that important contributions will be overlooked in the scope of the review, which can lead to an incomplete or fragmented understanding of the discipline.

This work adopts a methodology supported by Large Language Models (LLMs), utilizing these tools to facilitate the systematic and scalable exploration of scientific literature. This approach draws on traditional research methods, applying textual analysis to processed and manually annotated data [3]. LLMs are used as computational tools capable of processing natural language, identifying patterns, and assisting in the preliminary organization of content without replacing the critical analysis performed by experts. Their application works with the review process. They provide suggestions and mappings that human researchers continually evaluate, adjust, and validate. This approach is inspired by the principles of traditional reviews, incorporating LLMs as mechanisms to support the progressive generation of annotations on a manually audited dataset. By integrating LLMs into the workflow, the methodology seeks to enhance scalability and consistency while maintaining the rigor of scientific inquiry. In this sense, LLMs act as facilitators of discovery rather than as substitutes for human reflection, enhancing the capacity for synthesis without compromising the interpretive responsibility of the researcher. However, it is important to consider that LLMs can be vulnerable to membership inference attacks, where information about the training data can be leaked, including through fixed embedding models or via collision attacks on synthetic data generated by Generative Adversarial Networks (GANs). Furthermore, the ability to communicate covertly via cryptographically secure steganography using LLMs can impact Cyber Threat Intelligence (CTI) use cases involving stealth signaling or undetectable Command-and-Control (C&C) [4]–[7].

In the dynamic domain of cybersecurity, the exponential proliferation of data—from incident reports, intelligence feeds, forum communications, and vulnerability documentation—burdens analysts and researchers in the critical task of generating CTI. To mitigate this overload and extract actionable knowledge, Natural Language Processing (NLP) techniques and, most notably, LLMs represent significant advances in the treatment of textual information. Figure 1 illustrates the conceptual framework that guides this investigation, highlighting the interaction between the pillars, as well as data proliferation and the review process, in the context of cybersecurity. Thus, this article proposes to categorize, analyze, and synthesize the current state-of-the-art literature regarding the application of CTI, NLP, and LLMs as cyber defense resources. Our review process enabled us to identify various approaches and techniques employed in cybersecurity. We evaluated their effects, identified ongoing challenges, and noted potential future research opportunities to enhance cybersecurity solutions.

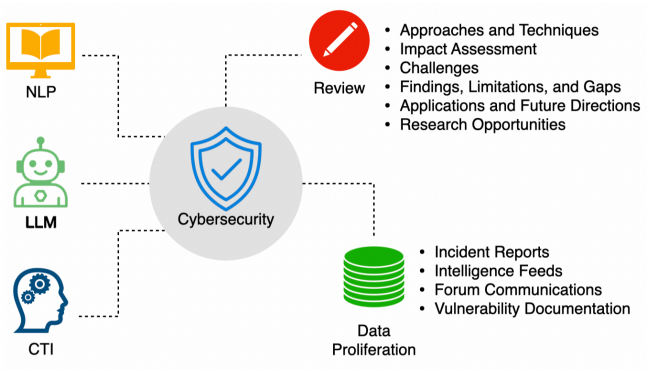

FIGURE 1. Conceptual framework for the application of LLMs, NLP, and CTI in the context of cybersecurity.

As a result of this review, this paper provides insight into the application of CTI, NLP, and LLMs in the cybersecurity domain [8], [9]. Our main contributions include categorizing the various existing approaches that employ these technologies to strengthen cybersecurity, analyzing the strengths, limitations, and challenges inherent in their application in threat environments, and identifying a set of emerging research opportunities that outline future directions for innovation in this field. This synthesis aims to guide researchers and practitioners in operationalizing LLMs and NLP in scalable, real-world CTI applications. Furthermore, multi-expert-based detection frameworks, initially designed to mitigate "phantom attacks" in driver assistance systems, offer a transferable approach to identifying visual deception in CTI lifecycles that process visual threat indicators [10].

The remainder of this paper is organized as follows. Section II introduces background concepts and surveys related work at the intersection of CTI, NLP, and LLMs. Section III describes the methodology adopted for literature collection and analysis. Section IV analyzes the applications of CTI, NLP, and LLMs in cybersecurity, categorizing the identified approaches. Section V discusses the main findings, including emerging trends, challenges, and research opportunities revealed by the literature synthesis. Finally, Section VI summarizes the key contributions and outlines future research directions.

A. OBJECTIVES AND RESEARCH QUESTIONS

The purpose of this work differs from what is typically seen in more traditional literature reviews. Instead of simply mapping what has already been done at the intersection of CTI, NLP, and LLMs, we attempt something more exploratory: identi-

fying even subtle signals of areas still understudied—spaces where there may be an impact disproportionate to the attention they are receiving. We use the metaphor of "searching for diamonds" here not as a frill, but as a way to describe this attempt to carefully unearth research potential hidden in plain sight.

To this end, we developed an approach that combines the use of LLMs with human curation. The idea is to enable a broader, more structured scan of the scientific literature—not only scalable, but also auditable and minimally biased. The process involves expert-guided prompt engineering, supported by language models, to collect, organize, and classify textual data taken from recognized cybersecurity sources. It is a kind of large-scale active listening. Three research questions (RQ) guided the study:

RQ1: What cross-domain research opportunities emerge from the integration of LLMs into CTI workflows?

RQ2: At what stages of the CTI lifecycle can LLMs offer functional, scalable, and reliable contributions?

RQ3: What are the architectural, methodological, and operational considerations for incorporating LLMs into future CTI systems?

From these questions, we raise a hypothesis that, while bold, seems reasonable: LLMs applied in a structured manner across various stages of the CTI cycle—from data collection to prioritization and alert dissemination—can enable a new generation of more modular, responsive, and context-aware architectures. This does not mean replacing human experts, but perhaps shifting the role LLMs play from auxiliary tools to more central elements of the process.

Our long-term goal, then, is to contribute to the collaborative design of an architecture in which these models do not remain on the periphery of CTI but become an integral part of the intelligence lifecycle. The work presented here aims to pave this path: to identify what is already working, pinpoint the main bottlenecks, and address the technical or conceptual obstacles that still need to be resolved before we move toward a more systematic and reliable approach.

II. BACKGROUND AND RELATED WORK

With the increasing popularity of cybersecurity, there is a growing demand for scalable and intelligent methods that can effectively handle the overwhelming volume of data and knowledge characteristic of this domain. This section reviews prior work at the intersection of NLP, LLMs, and cybersecurity, highlighting its key contributions to CTI. Table 1 outlines the dimensions that interconnect CTI, cybersecurity, and operational workflows, reflecting the essence of the challenges and solutions explored in this section.

NLP techniques have increasingly been adopted in cybersecurity to extract structure from unstructured data sources, including threat reports, malware descriptions, and dark web forums. These techniques play a relevant role in transforming raw data — such as incident reports, intelligence feeds, forum discussions, and vulnerability documentation — into structured, actionable insights. Concurrently, LLMs have extended traditional NLP pipelines by enabling contextual reasoning, semantic abstraction, and zero-shot generalization. In CTI-focused research, NLP methods have supported the extraction of indicators, text classification, and the summarization of threat narratives. Despite this progress, most existing studies operate in isolated settings and focus on constrained use cases, synthetic datasets, or demonstrative prototypes, limiting their integration into broader CTI workflows.

This fragmentation underscores the need for a more unified perspective—one that not only maps technical capabilities but also considers their role within strategic cybersecurity contexts, where interpretability, scalability, and operational relevance are valuable. To outline this landscape, we explore key contributions across NLP methods related to cybersecurity, emerging uses of LLMs in security contexts, and approaches that can be directly applied or adapted for CTI workflows. For each area, we analyze model architectures, task definitions, evaluation methods, and constraints as outlined by the authors. Many studies may not be directly identified as CTI contributions, but they offer valuable elements for CTI workflows, including robust representation learning, privacy-preserving inference, and semantic parsing of threat intelligence documents. When integrated, these building blocks enable the development of adaptive CTI lifecycles suited to dynamic threat environments.

The reviewed literature highlights advances in binary decompilation using language models, log anonymization, and robust classification under obfuscation. However, these contributions remain siloed across distinct research efforts and lack a unifying framework. To address this disconnect, our work proposes a modular lifecycle that aggregates these dispersed techniques, applies human-like reasoning over a structured cybersecurity corpus, and identifies strategic research opportunities at the intersection of CTI, NLP, and LLMs.

A. NLP FOR CYBERSECURITY

Recent progress in NLP, particularly advancements fueled by LLMs and Deep Learning (DL) architectures, has considerably improved the analytical abilities of cybersecurity systems. These innovations have led to the creation of new tools for extracting semantic structures from binaries, deciphering obfuscated content, and ensuring data confidentiality in hostile environments. While often not directly associated with a CTI framework, the methods are essential for CTI operations. This includes malware reverse engineering, provenance-aware code attribution, and scalable knowledge extraction from open-source threat intelligence sources. By enabling more nuanced interpretation, transformation, and safeguarding of complex data types, these methods provide vital components for strong, intelligence-driven cybersecurity processes. Figure 2 illustrates the thematic trajectories emerging from NLP research in cybersecurity, highlighting how distinct streams—from textual data analysis to adversarial robustness—converge into operational challenges and security risks relevant to CTI.

A growing body of research investigates Transformer-

TABLE 1. Key Technical Dimensions Across CTI, Cybersecurity, and Operational Workflows

Technical Dimension	Application in CTI	Application in Cybersecurity	Application in Workflows
NLP	Structuring threat intelligence from unstructured text	Extracting security-relevant insights from logs and reports	Automating understanding and integration of text in workflows
LLMs	Contextual prediction and reasoning for threat detection	Generalized modeling of cyber phenomena	Enhancing decision support through semantic representations
Transformers	Modeling attacker behavior and tactics	Foundation for modern NLP in cyber research	Architectures supporting scalable analysis tasks
Semantic Extraction	Summarizing and labeling threat indicators	Representing patterns in cyber telemetry	Optimizing workflow logic based on semantic roles
Reverse Engineering	Unpacking binaries for malware insights	Identifying vulnerabilities in compiled code	Embedding reverse analysis into toolchains
Instruction Embedding	Representing low-level instructions semantically	Comparing binary-level behaviors	Enhancing binary analysis tools in automated chains
Privacy-preserving Inference	Safeguarding intelligence in adversarial settings	Protecting data in LLM/NLP inference	Securing collaborative model execution
Differential Privacy	Preventing data leakage from training logs	Privacy-preserving model learning	Compliance-aware model deployment
Secure MPC	Multi-party secure model execution	Privacy in distributed threat modeling	Integrating encrypted inference in lifecycles
Adversarial Robustness	Detecting manipulation of threat indicators	Hardening models against adversarial inputs	Ensuring resilience in automated systems
Model Poisoning	Manipulating LLM behavior during training	Compromising integrity of ML-based defenses	Mitigating training-phase attacks in lifecycles

based frameworks for recovering semantics from stripped binaries, a common skill in reverse engineering. Kim et al. [11] present AsmDepictor, a Transformer-based sequence-to-sequence model that derives function names directly from assembly code. Their approach, which features domain-specific enhancements such as per-layer positional embeddings and a Jaccard-based evaluation metric, demonstrates notable performance compared to existing benchmarks, indicating its potential for behavior labeling and identifying threat actor tactics in CTI. Likewise, Jin et al. [12] unveil SymLM, a context-aware embedding system that combines instruction semantics and calling contexts, improved by a customized function name preprocessing lifecycle. Their model demonstrates excellent generalization across various architectures and obfuscations, which are important for practical CTI applications in diverse binary environments. DIRTY [13], a Transformer-based system, enhances existing methods by retrieving variable types and names from decompiled code with an accuracy exceeding 66% for name recovery. While not explicitly targeted at CTI, DIRTY's multi-task framework—integrating code and memory layout encodings—supports detailed semantic interpretation of binary artifacts, a valuable aspect in malware reverse engineering and knowledge graph creation for CTI platforms. Concurrent initiatives in instruction embedding focus on improving foundational code comprehension. PalmTree [14], inspired by BERT and designed for assembly code, utilizes self-supervised goals (like context prediction and def-use relations) to develop rich, transferable embeddings that work across different compilers and ISAs. This advancement supports several downstream tasks, including binary similarity and control-flow inference, which are essential for clustering, lineage tracking, and behavior correlation in CTI systems. In a similar vein, INNEREYE [15] employs neural machine translation methods to conduct cross-architecture similarity analysis, surpassing symbolic tools in both precision and speed. Its effectiveness in identifying reused cryptographic code and network parsers across various architectures clearly aligns with CTI requirements for analyzing embedded malware and tracking threat campaigns.

Building on this momentum, recent research has introduced privacy-preserving frameworks for Deep Neural Network (DNN) and Language Model (LM) inference, addressing critical challenges for secure deployment in collaborative or sensitive environments—an increasingly significant issue in CTI applications. Akimoto et al. [16] propose Privformer, a system enabling complete Transformer inference through a three-party secure Multi-Party Computation (MPC) protocol based on the honest-majority assumption. This framework substitutes the computationally intensive softmax with ReLU using the Performer approximation, effectively lowering the complexity of attention mechanisms while maintaining functionality. It also incorporates masked attention strategies and a novel square-root inverse protocol to ensure data privacy and operational feasibility. Privformer enables 64-token inference in 19 minutes over a LAN with 4.64 GB of data traffic, demonstrating its practicality for encrypted inference in CTI lifecycles that manage sensitive intelligence artifacts. Additionally, Du et al. [17] introduce DP-Forward, a forward-pass noise injection technique based on the analytic Matrix Gaussian Mechanism (aMGM), which provides formal Local Differential Privacy (LDP) assurances at the sequence level. This new approach differs from Differentially Private Stochastic Gradient Descent (DP-SGD) by enforcing privacy restrictions directly on hidden representations, thereby reducing risks such as embedding inversion and sensitive attribute inference, while also enhancing training efficiency and accuracy by up to 7.7 percentage points compared to DP-SGD. This level of granular control over sequence-level privacy is especially

beneficial in CTI tasks where both training and inference involve proprietary or classified threat data. Concurrently, Huang et al. [18] present Cheetah, a high-performance two-Party Computation (2PC) system for secure DNN inference, utilizing SIMD-free homomorphic encryption for linear layers and VOLE-based oblivious transfer for non-linear layers. With support for large-scale architectures like ResNet50, Cheetah achieves a 5.6× speedup and a 12.9× reduction in communication compared to CrypTFlow2, completing full inference in under 2.5 minutes. While none of these systems were explicitly designed for CTI, their strong privacy guarantees, scalability, and low-latency performance create a strong foundation for CTI-aligned workflows. In particular, they facilitate secure collaboration across organizational boundaries, encrypted inference over sensitive indicators, and private fine-tuning of LLMs with proprietary intelligence—all essential for integrating Artificial Intelligence (AI) within the trust and compliance frameworks required by modern cybersecurity ecosystems.

In addition to advancements in privacy-preserving inference, another area of research has focused on the adversarial vulnerabilities and defenses of NLP models. This research highlights integrity threats and resilience strategies for LLM-based lifecycles in the context of CTI. Pei and Yao [19] present a two-stage framework for creating content-preserving yet meaning-inverting (CPSF) adversarial examples that reveal NLP models' insensitivity to semantic changes. By combining semantic flipping through either rule-based or neural transformations with black-box perturbation methods, they produce coherent adversarial texts that effectively mislead classifiers, such as fastText and CNN-RNN, without sacrificing human readability. These attacks demonstrate significant transferability to commercial NLP APIs, revealing a substantial vulnerability that could be exploited to manipulate CTI systems interpreting threat narratives. Similarly, Boucher et al. [20] investigate subtle encoding-level perturbations, such as invisible Unicode characters and bidirectional reordering, which alter model behavior without changing visible content. These tactics exhibit high transferability across various tasks and models, sometimes diminishing accuracy by over 90% and even evading indexing in commercial search engines. The implications for CTI are significant, as these strategies allow malicious actors to avoid detection in open-source feeds, corrupt datasets used for LLM fine-tuning, or covertly disrupt downstream inference. Transitioning from offensive strategies to defensive measures, Anjum et al. [21] introduce a sequence-to-sequence attention-based model for automatically anonymizing clinical texts, redefining anonymization as a generative transformation task rather than merely a token classification task. This architecture, achieving state-of-the-art recall on the i2b2 benchmark, could be adapted for CTI scenarios that necessitate the automatic redaction of sensitive threat indicators from logs or intelligence reports. Additionally, Schuster et al. [22] reveal how word embeddings developed from public datasets are vulnerable to subtle poisoning attacks that manipulate co-occurrence statistics to distort semantic distances. These attacks, formulated as optimization problems over embedding spaces, can impair downstream tasks such as Named Entity Recognition (NER) and retrieval, circumventing standard defenses like perplexity filtering. This poses a significant risk to CTI workflows that rely on pre-trained embeddings or public intelligence datasets, as compromised data could jeopardize model integrity without quick detection. Finally, Li et al. [23] propose TEXTSHIELD, a robust classification system aimed at detecting toxic content in Chinese. This system merges adversarial machine translation with multimodal glyph, semantic, and phonetic embeddings, making it resistant to adaptive and transfer attacks. TEXTSHIELD enhances classification accuracy by as much as 45% on real-world adversarial inputs, thus offering a setup that could be useful in CTI workflows focused on Chinese-language threat sources or social media monitoring. These studies highlight the vulnerability of NLP models in adversarial settings and provide concrete approaches to enhance their robustness. Their findings are particularly relevant for CTI systems functioning in hostile, multilingual, and data-contested environments, whether aimed at reinforcing LLM-based analysis pipelines against covert manipulations or ensuring secure handling of intelligence data amid increasingly sophisticated threats.

Beyond adversarial attacks and privacy-protecting inference, research has increasingly focused on semantics-aware methods for detecting, transforming, or securing natural language and speech functionalities essential to CTI processes. These processes involve covert communication, voice-driven interfaces, and tracking the origin of generated content. Yuan et al. [24] introduce Cantreader, a dual-stage embedding-based system that identifies and interprets "dark jargon" by assessing semantic divergence across databases from underground forums, reputable sources, and encyclopedic entries. Cantreader achieves a 91% precision rate in identifying 3,462 jargon terms and 478 previously unrecognized blacklisted expressions, thereby enabling the automated detection of covert threat language—a promising tool for CTI efforts that rely on the early identification of evolving terminology in cybercrime communities. In the auditory space, Guo et al. [25] propose SkillExplorer, a grammar-driven framework for analyzing large-scale behavior of voice assistant skills, which identifies systemic privacy breaches and misconfigurations across over 30,000 Alexa and Google Assistant apps. Their technique, rooted in simulated multi-round interactions and grammar-based input generation, can be adapted for auditing conversational CTI interfaces, particularly to identify unauthorized data collection or covert surveillance patterns in LLM-integrated virtual agents. On the topic of voice anonymization, Deng et al. [26] introduce V-CLOAK, a real-time anonymizing tool that employs Wave-U-Net, utilizing voiceprint modulation and adaptive perturbation control. This approach achieves top-tier anonymization metrics (EER = 46.10%) while maintaining speech clarity (WER = 7.65%). Although aimed at preserving speaker privacy, its ability to anonymize voice reports while retaining semantic content

is immediately applicable in CTI contexts that require secure verbal communication or protected analyst collaboration via voice channels. Meanwhile, Abdelnabi and Fritz [27] propose the Adversarial Watermarking Transformer (AWT), an innovative sequence-to-sequence model for embedding resilient multi-bit watermarks in natural language using learned word substitutions. AWT simultaneously optimizes for stealth, message recoverability, and semantic integrity, achieving 97.04% decoding accuracy and resistance to removal attempts. This approach can be directly applied within CTI workflows to monitor the origin of LLM-generated intelligence artifacts, facilitating transparent auditing without compromising the authenticity of the content. These studies showcase diverse semantics-aware protections—such as lexical detection, audio anonymization, and linguistic watermarking—that enhance CTI capabilities beyond traditional text classification and entity extraction. They provide effective methods for securing multimodal threat intelligence processes, improving the dependability, interpretability, and accountability of CTI systems in adversarial or privacy-sensitive environments.

Alongside these semantics-aware protections, another research avenue focuses on program analysis and neuro-symbolic reasoning methods to recover, certify, or approximate Machine Learning (ML) behavior under constrained or opaque conditions. These situations are particularly significant for CTI applications involving embedded AI, obfuscated malware, or adversarially altered logic. Wu et al. [28] introduces DnD, a system that is both compiler- and ISA-agnostic and can decompile DNNs from stripped binaries. By employing symbolic execution, loop analysis, and Abstract Syntax Tree (AST) matching, DnD reconstructs comprehensive ONNX models that encompass operator types and parameters across various architectures, including Thumb, AArch64, and x86-64. DnD facilitates white-box attacks on embedded systems through its ability to recover most evaluated models perfectly. This showcases its utility in reverse-engineering AI-enhanced malware and extracting neural components from firmware-based threat artifacts in CTI. To tackle robustness within perception lifecycles, Yuan et al. [29] present GCERT, a framework that certifies the resilience of neural networks against semantic image mutations by modeling transformations as linear traversals within generative latent spaces. By applying structural constraints on the generative model—ensuring independence and continuity—GCERT supports complete, incomplete, and quantitative certification across tasks such as classification and face recognition. These skills are especially relevant for assessing the reliability of vision-based threat indicators when faced with real-world variations, such as lighting or stylization changes. Lastly, Shen et al. [30] introduce NEUEX, a neuro-symbolic execution engine that builds on classical symbolic execution by incorporating learned neural constraints. NEUEX dynamically generates neural approximations when symbolic solvers fail to cover paths obstructed by opaque logic or external libraries. It can uncover vulnerabilities that are overlooked by standard tools, such as KLEE. It performs exceptionally well in analyzing loop-intensive or partially observable code scenarios, which are frequently found in obfuscated malware samples or evasive threat payloads. These systems, originally not designed for CTI, establish a foundation for reasoning with AI-integrated binaries, validating ML behavior amid input variations, and interpreting complex logic in advanced threat artifacts. Collectively, they enhance the analytical capabilities of LLM-augmented CTI lifecycles, offering interpretable, certifiable, and reverse-engineerable assessments of adversarial and AI-driven threat vectors.

Despite technical advances in NLP methods for cybersecurity, the reviewed studies generally focus on isolated capabilities and do not integrate into broader CTI workflows. Moreover, few works evaluate these methods in real-world operational settings or assess their interaction with other components of an intelligence lifecycle. This gap highlights the need for a more unified approach that not only leverages the potential of NLP to process large amounts of data but also integrates these capabilities into CTI lifecycles in a scalable and operational manner. This reinforces the need for systematic, reproducible frameworks that can consolidate and operationalize such capabilities at scale.

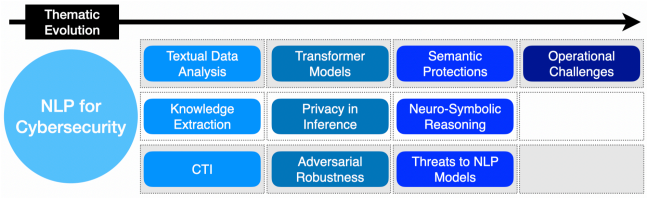

FIGURE 2. Application of NLP in Cybersecurity.

B. EMERGING USE CASES OF LLMS

This subsection compiles rapidly advancing research initially outside cybersecurity that offers interesting capabilities increasingly promising to CTI enhanced by LLMs. As LLMs become increasingly valuable in threat analysis and incident response, techniques from adversarial NLP, privacy-preserving inference, semantic analysis, and secure neural architectures have demonstrated significant applicability in CTI. We categorize this literature into four key areas: (i) adversarial robustness and evasion, focusing on LLM manipulation in classification pipelines; (ii) privacy leakage and model auditing, addressing risks of memorization and IP theft; (iii) semantic extraction and behavioral labeling, organizing and examining user or adversary-generated content for threat assessment; and (iv) secure neural architectures and program analysis, highlighting GNN-based inference and encrypted execution. Collectively, these studies show that while LLMs enhance CTI analytical capabilities, they also introduce novel risks and attack vectors, necessitating the integration of protective and analytical strategies for their secure use in hostile environments.

Recent advancements in attacks on NLP and LLM-based systems have exposed vulnerabilities with profound implications for CTI lifecycles that ingest public, potentially manipulated, or adversarial inputs. [31] conduct an evaluation of six deepfake text detectors across multiple LLM-generated datasets, revealing that their performance can collapse under distributional shifts, up to 99% F1 degradation, highlighting a severe gap in robustness. They further introduce DFTFooler, a query-free black-box attack that achieves a 91.3% evasion rate through semantically guided token substitutions, preserving fluency and human interpretability. Similarly, [32] proposes TEXTBUGGER, an attack framework that perturbs input texts at character and word levels to evade detection by major commercial NLP services. With success rates nearing 100% and over 94% human interpretability retained, TEXTBUGGER demonstrates how subtle, undetectable input perturbations can compromise LLM-based classifiers, posing direct risks to sentiment- and toxicity-filtering stages in CTI lifecycles. Addressing a more insidious attack vector, [33] unveils the Zero-Width (ZeW) attack, which inserts invisible Unicode characters to disrupt tokenization without affecting human readability. Evaluated across 12 commercial APIs, ZeW degrades NLP model performance in 11, underscoring the fragility of input preprocessing stages frequently outsourced in CTI systems. Ditto introduces a more covert manipulation [34], a model hijacking method that poisons the training data of text generation models (e.g., summarizers or translators), covertly embedding classification behaviors. With attack success rates exceeding 93% and minimal degradation to utility, Ditto reveals how fine-tuned models in CTI applications might silently leak or alter threat narratives. In highly constrained settings, [35] presents RamBoAttack, a decision-based black-box attack that integrates gradient estimation and localized perturbations to outperform prior methods in efficiency and success, particularly relevant to red-teaming and adversarial testing of externally deployed CTI classifiers. Lastly, [36] introduces GEESOLVER, a semi-supervised CAPTCHA solver that leverages masked autoencoders to break complex visual CAPTCHAs with over 90% success using minimal labeled data. While initially targeting security challenges beyond CTI, its architecture generalizes to pattern recognition problems within CTI, such as decoding obfuscated threat artifacts or extracting adversarially hidden indicators. Collectively, these studies illuminate a critical blind spot in current CTI practices: adversarial resilience. The success of techniques in bypassing commercial NLP tools and exploiting input preprocessing highlights the urgent need to strengthen CTI systems against input manipulations and model misuse, particularly as LLMs become increasingly crucial in threat analysis. The susceptibility of LLMs to such attacks underscores the complexity of relying on automated systems to process the vast and malicious volume of cybersecurity data, presenting a central challenge.

Building on the previously mentioned vulnerabilities, an associated area of research explores privacy leaks, model extraction, and the safeguarding of intellectual property within neural networks. These issues are becoming increasingly significant in CTI workflows that utilize or optimize LLMs on sensitive information. The work by Carlini et al. [37] presents the exposure metric to measure unintended memorization in generative models. Their findings indicate that character-level LLMs can disclose secrets from training datasets, such as credit card numbers, even without overfitting, unless differentially private training methods (e.g., DP-SGD) are utilized. Their investigation into academic models and commercial systems, such as Google Smart Compose, reveals that conventional regularization strategies (e.g., dropout) fall short, underscoring the importance of systematic privacy audits in LLM-driven CTI processes that handle sensitive reports or incident information. Looking further into privacy issues, [38] proposes GAN-Leaks, a comprehensive framework for membership inference attacks targeting GANs and VAEs. Their work demonstrates that high AUC-ROC scores (over 0.95) can be achieved across various datasets and model types, even in limited access environments, raising concerns regarding CTI systems that generate synthetic threat data through generative models—data that might inadvertently reveal whether specific incidents were included in the training set. From a forensic perspective, [39] introduces DEEPJUDGE, a post hoc model auditing framework that surpasses traditional watermarking and fingerprinting techniques by calculating multi-level similarity metrics (such as neuron and layer activation distances) to identify unauthorized model usage. This method facilitates tracking LLM provenance even in black-box settings, which is essential for confirming the source and integrity of CTI models shared across different agencies or partners. In relation to model extraction, [40] showcases both task- and fidelity-oriented attacks that reconstruct neural network functions based solely on query access. They notably achieve the first precise weight recovery of ReLU networks under black-box situations, providing a concrete methodology for reverse engineering or verifying publicly accessible LLM-based CTI systems via APIs. Likewise, [41] establishes an extraction lifecycle designed for Deep Reinforcement Learning (DRL) agents, employing adversarial imitation to mimic agent behavior with a 97% fidelity rate—demonstrating the potential risks to DRL components within CTI, such as autonomous threat triage or incident response protocols. To protect intellectual property in these adversarial circumstances, [42] presents a robust DNN watermarking scheme that utilizes trigger-based queries, remaining effective even against model pruning and inversion attacks. Their approach enables verifiable ownership claims in white-box and black-box contexts, providing a practical means to ensure model provenance and prevent misuse in federated or collaborative CTI settings. Together, these findings contribute to an essential set of capabilities for LLM-powered CTI systems: the ability to assess privacy exposure, detect unauthorized usage, and safeguard the integrity of models operating in dynamic, distributed, and adversarial environments. As CTI increasingly relies on optimized LLMs and generative models for generating, sharing, and analyzing threat intelligence,

these methods prioritize privacy and provenance, providing a critical layer of security and accountability. These surveys reflect the growing concerns already highlighted regarding public trust and investigative transparency, particularly when critical cybersecurity systems rely on sensitive data and advanced technologies. Figure 3 summarizes the role of LLMs in cybersecurity, mapping their contributions—from enhancing threat analysis to securing semantic lifecycles.

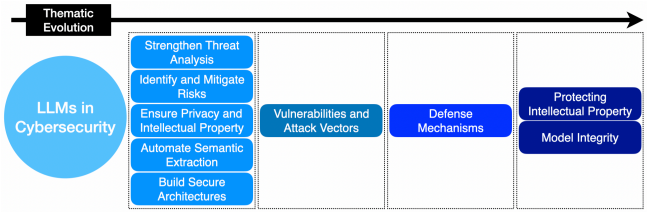

FIGURE 3. Emerging Applications of LLM in Cybersecurity.

Extending this perspective, another emerging research direction focuses on semantic extraction and behavioral classification techniques applied to large-scale, user-generated, or adversary-generated content, offering methodologies that directly enhance CTI workflows. Harkous et al. [43] introduce Hark, a modular, end-to-end lifecycle founded on five task-specific T5-11B models designed to extract and structure privacy-related information from unstructured text. When applied to 626 million Google Play reviews, Hark achieves an impressive accuracy rate of 96% in abstractive issue generation and identifies over 300 high-level semantic themes, supported by clustering and emotion-aware annotation modules. Although not initially intended for CTI, Hark's architecture demonstrates the required scalability and adaptability to convert unstructured reports into actionable intelligence, such as user-submitted threat indicators or incident descriptions. Similarly, Prasad et al. [44] present SnorCall, a behavioral analysis system for robocall campaigns that integrates automated speech transcription, weak supervision via Snorkel, and semantic labeling to classify over 230,000 scam-related calls over a two-year period. The framework reveals infrastructural reuse and thematic patterns, including impersonation and political manipulation. It creates interpretable taxonomies, illustrating how such lifecycles can be repurposed to analyze large volumes of adversarial communications or phishing voice attacks within CTI environments. Complementing these efforts, Wang et al. [45] propose Sacabuche, a semantic filtering and validation system that detects manipulation within autocomplete suggestions—an overlooked yet potent attack vector exploited for search engine poisoning and social engineering. By combining linguistic similarity metrics with search result analysis, Sacabuche achieves over 96% precision and recall on 114 million trigger–suggestion pairs, identifying hundreds of thousands of instances of adversarial manipulation. The system's language-aware triage and validation methodology offers a transferable blueprint for detecting coordinated influence operations or semantic poisoning attempts within Open-Source Intelligence (OSINT) feeds utilized by CTI platforms. Collectively, these works exemplify how large-scale, LLM-compatible semantic analysis systems—despite not being initially designed for cybersecurity—can be adapted to extract structured insights, uncover behavioral patterns, and monitor evolving threats across noisy, untrusted, or adversarial data streams. Their integration into CTI lifecycles holds the promise of enhancing early warning, infrastructure correlation, and adversarial content detection with high accuracy and operational scalability.

Recent studies on secure neural architectures and sophisticated program analysis methods show great promise for improving CTI lifecycles, particularly in areas such as anomaly detection, memory-aware reverse engineering, and privacy-preserving inference. King and Hohman [46] present EULER, a scalable framework for temporal link prediction that combines Graph Neural Networks (GNNs) with recurrent sequence models to identify lateral movement in enterprise networks. In evaluations using the 1.6 B-event LANL dataset, EULER demonstrates a positive detection rate of over 90% and superior precision compared to GNN and rule-based methods. Its ability to separate spatial from temporal encoding makes it especially suited for CTI workflows that necessitate high-fidelity detection of stealthy, evolving threats. Nonetheless, as Shen et al. [47] indicate, GNN-based systems can be susceptible to model-stealing attacks, even with limited knowledge. Their two-stage attack reconstructs inaccessible graph structures and trains precise surrogate models, exposing vulnerabilities in graph-based CTI systems used for malware infrastructure clustering or adversarial network mapping. In a different approach, Lyu et al. [48] introduce GOSHAWK, a framework for detecting memory corruption that utilizes NLP-based classification alongside symbolic execution to identify unique memory behaviors across complex codebases. By uncovering 92 new bugs—including use-after-free and double-free errors—in kernel and Internet of Things (IoT) software, GOSHAWK demonstrates its relevance in CTI contexts where memory manipulation is employed to evade detection or ensure persistence. Similarly, She et al. [49] propose NEUTAINT, a neural taint analysis framework that substitutes traditional rule-based tracking with gradient-guided influence models, resulting in improved accuracy, broader coverage, and significantly reduced runtime for fuzzing and Common Vulnerabilities and Exposures (CVE) discovery. With an average increase of 61% in edge coverage and effective real-world CVE detection, NEUTAINT lays a strong foundation for taint-aware malware analysis and vulnerability discovery within CTI lifecycles enhanced by LLMs. Lastly, addressing confidentiality in CTI applications, Jang et al. [50] create MatHEGRU, a GRU-based sequence model that enables end-to-end encrypted inference using the MatHEAAN homomorphic encryption technique. MatHEGRU achieves near-plaintext accuracy (e.g., 94.2% on MNIST) while ensuring the privacy of both the model and data, facilitating secure LLM-driven sequence analysis, such as encrypted behavioral modeling of malware or genomic

data in cyber-biosecurity scenarios. These initiatives indicate an advancing trend in CTI system design: the integration of secure model architectures, privacy-preserving analytics, and cutting-edge program analysis. While originally devised in related fields, their application within CTI introduces tangible improvements in robustness, interpretability, and confidentiality—key features for mission-critical, adversary-aware settings.

Together, these works expand the frontier of LLM applications in security contexts, providing concrete tools for enhancing the robustness, traceability, and privacy of CTI systems. Their integration into threat intelligence lifecycles opens new directions for secure automation, adversarial resilience, and semantic enrichment in real-world intelligence workflows.

C. DIRECT OR POTENTIAL APPLICATIONS OF LLMS IN CTI

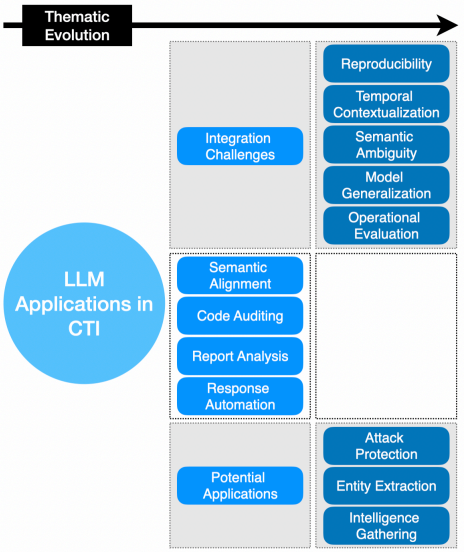

FIGURE 4. LLM Applications in CTI

The security of LMs and ML systems is becoming increasingly essential in adversarial contexts, such as CTI, where proprietary models handle sensitive information and inform defensive strategies. Asokan [51] analyzes model extraction attacks via inference APIs, stressing how adversaries can mirror commercial models using adaptive queries. To counter this, Model Ownership Resolution (MOR) methods, such as watermarking and fingerprinting, are explored; however, the author notes that combining these with privacy-preserving techniques may compromise robustness. In a related study, Lukas et al. [52] formalize and assess three black-box attack methods aimed at leaking Personally Identifiable Information (PII) from LMs fine-tuned on private datasets. Their investigation reveals significant leakage (3–8%) even in differentially private models ($epsilon = 8$), highlighting the shortcomings of current differential privacy implementations in real-world adversarial conditions. These privacy issues pose risks to CTI systems that manage classified or sensitive documents, where membership inference or contextual reconstruction could compromise operational confidentiality. In a more aggressive context, Zhang et al. [53] present TROJANLM, an attack that embeds logic-based triggers into LMs, leading to targeted misclassifications with remarkable stealth and effectiveness ($ASR \geq 90\%$), potentially jeopardizing CTI workflows by subtly altering NLP-driven evaluations. To counter this risk, Azizi et al. [54] introduce T-Miner, a detection framework that functions without prior knowledge of clean inputs and achieves high accuracy and resilience across various architectures through a perturbation-based probing method. Although each study addresses different attack vectors—extraction, leakage, poisoning, and detection—they all underscore an urgent need for comprehensive security models that harmonize privacy, integrity, and provenance within ML systems. For CTI, where maintaining model integrity is crucial for situational awareness and response efficacy, these findings advocate for cohesive defenses that are both robust and capable of adapting to diverse threat landscapes.

In response to the need for defenses, initiatives have introduced frameworks integrated with LLMs that enhance CTI workflows through improved vulnerability modeling, code reasoning, and threat automation. Qin et al. [55] propose the Contextual Entity Alignment Model (CEAM), a domain-specific entity alignment model based on GNNs, which effectively addresses inconsistencies across heterogeneous vulnerability repositories such as the National Vulnerability Database (NVD), ICS-CERT, and SecurityFocus. By incorporating asymmetric aggregation and partitioned attention mechanisms, CEAM attains an F1 score of 84.3%, thereby facilitating the reliable correlation of fragmented security intelligence and enhancing the consistency and comprehensiveness of CTI databases. Conversely, Pearce et al. [56] disclose systemic vulnerabilities in LLM- assisted software development by evaluating code completions from GitHub Copilot, wherein over 39% contain Common Weakness Enumeration (CWE) flaws, including command injection and use- after-free vulnerabilities. These findings underscore a paradox: while LLMs expedite CTI coding tasks, they may inadvertently perpetuate insecure patterns derived from training data, highlighting the necessity for integrated static analysis and safety constraints. Additionally, Wang et al. [57] introduce iRuler, a formal framework designed for detecting logic flaws within IoT automation platforms, such as IFTTT. Through Satisfiability Modulo Theories (SMT)-based verification and NLP-assisted flow inference, iRuler reveals six inter-rule vulnerabilities across more than 300,000 real-world applets. This illustrates that CTI systems incorporating intelligent automation must account for logic-level misconfigurations and chaining behaviors that could compromise response mechanisms. Finally, Quiring et al. [58] connect the adversarial paradigms of ML and Digital Watermarking (DW), demonstrating that evasion, extraction, and oracle attacks across these domains can be unified under equivalent optimization formulations. This conceptual unification enables the adaptation of defense strategies, such as boundary detectors and

classifier diversity, across domains, making it a valuable tactic for safeguarding LLMs vulnerable to query-based probing in CTI environments. Collectively, these contributions demonstrate significant technical advancements and highlight the critical importance of domain-aware, model-centric, and cross-disciplinary approaches in strengthening CTI lifecycles against diverse and evolving threat vectors.

Extending this interdisciplinary momentum, other researchers have also advanced forensic and attribution techniques that enhance CTI lifecycles by targeting cybercrime infrastructures through multimodal analysis. This analysis includes financial transactions, image-based identity traces, blockchain activities, and steganographic artifacts. Gómez et al. [59] introduce a back-and-forth exploration, a transaction tracing approach for Bitcoin that simultaneously examines both forward and backward flows, significantly improving the discovery of covert financial ties among malware campaigns. By leveraging large-scale address tagging and ML operations, the technique identifies hidden C&C nodes and operational reuse across thirty malware families, uncovering thirteen previously unknown C&C addresses and achieving a 93% success rate in relational inference. Similarly, Su et al. [60] propose DEFIER, a DL framework that classifies Ethereum exploit transactions across various stages of attack, utilizing sequence labeling over execution trace graphs. When applied to 2.35 million transactions, DEFIER detects over 475,475 malicious entries, including within seventy-five zero-day decentralized applications (DApps)—highlighting the forensic potential of blockchain technology in CTI, particularly for decentralized threat infrastructures. Complementing these financial and protocol-level perspectives, Wang et al. [61] introduce a photographic fingerprinting method for Sybil detection in darknet marketplaces, wherein vendors are re-identified through deep visual features extracted from product images, surpassing stylometric baselines and exposing 738 fraudulent vendor pairs. This image-based attribution method reflects the potential of visual modalities—often underestimated in CTI—to be operationalized for identity resolution in anonymized ecosystems. Further emphasizing the capability of relational inference, Na et al. [62] construct a graph-based knowledge model that links identifiers from forty million dark web pages and Bitcoin transactions to actors affiliated with The Shadow Brokers, revealing both direct and inferred connections to hacker forums and illicit services. Although semi-automated, their framework lays the foundation for more scalable entity resolution using ML technologies. Lastly, Chapman [63] addresses the detection of covert data flows through the Structural Anomaly Detector for Hidden Underlying Graphics (SAD THUG), a structural anomaly detector for steganographic malware within image containers. By modeling benign JPEG and PNG structures with finite-state automata, SAD THUG achieves a true positive rate of over 99% across eleven malware families, without relying on content inspection, which is crucial for identifying C&C channels that evade conventional detection methods. Collectively, these contributions illustrate that effective CTI cannot rely solely on textual or code-based signals; instead, it must encompass a comprehensive spectrum of modalities—including financial, visual, protocol, and structural dimensions—to detect, attribute, and neutralize threat actor infrastructure within an increasingly polymorphic threat landscape.

In parallel with advances in modeling, automation, and attribution, another critical dimension of CTI lies in understanding the operational utility, limitations, and structural behaviors of intelligence sources. Bouwman et al. [64] offer the first evaluation of large-scale, open-source CTI sharing through their analysis of the COVID-19 Cyber Threat Coalition (CTC), a volunteer-driven initiative. Despite involving over 4,000 participants, results show that only 2.6% of blocklisted domains were COVID-specific, with significant redundancy and mitigation lag largely attributed to an over-reliance on VirusTotal heuristics. Nevertheless, the CTC added value for novel, pandemic-specific threats, where preexisting mitigation mechanisms were slower to act. This highlights a recurring tension between openness, scalability, and quality assurance in peer-sourced CTI. Complementing this, Bouwman et al. [65] conduct the first comparative study of paid threat intelligence (PTI) services versus Open Threat Intelligence (OTI). Their findings reveal that, while overlap between vendors was minimal (2.5–4.0%), even when tracking the same actors, PTI was valued for its contextual richness, trustworthiness, and perceived actionability, rather than for its measurable accuracy or coverage. This diverges from assumptions in CTI research that emphasize quantitative metrics over practitioner heuristics, revealing a mismatch between academic frameworks and operational realities. Addressing the broader ecosystem, Li et al. [66] propose a systematic evaluation methodology encompassing six key metrics—volume, differential and exclusive contribution, latency, accuracy, and coverage—applied to 55 real-world indicator feeds. Their analysis reveals low feed overlap ($typically < 10\%$), weak correlation between volume and quality, and inconsistent performance across categories and time windows. Notably, over 90% of indicators were unique to a single feed, and operational value varied by use case. Together, these works challenge the assumptions of data redundancy and coverage uniformity in CTI and advocate for more nuanced, goal-specific feed selection strategies. By quantifying structural trade-offs across open and commercial sources, and by aligning evaluation methods with practical utility, these studies advance CTI from a data aggregation problem to a decision-support discipline, where trust, context, and operational alignment are as critical as data fidelity or volume.

Recent advancements have built upon the necessity for domain-aware, context-specific intelligence by introducing specialized frameworks that address the unique challenges associated with CTI in domains often overlooked by conventional models, such as mobile communications, industrial systems, underground ecosystems, and graph-structured data. Chen and Ruddle [67] enhance the Bhadra framework

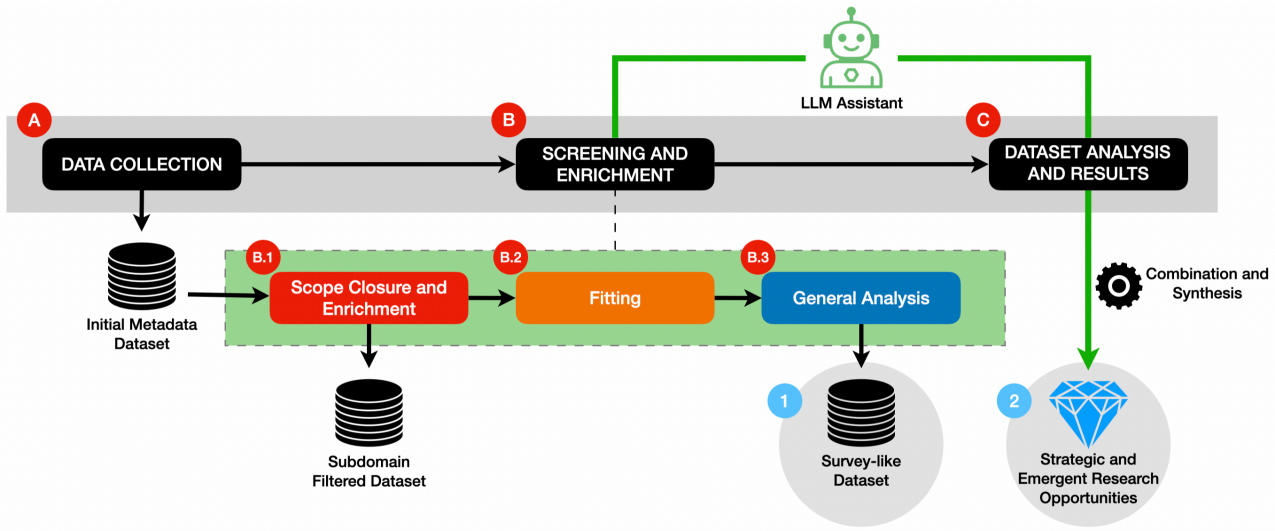

FIGURE 5. **Modular LLM-driven architecture for generating a survey-like dataset, and synthesizing strategic and emergent research opportunities.**

to facilitate structured threat modeling within mobile networks (2G–4G). This is accomplished through a modular toolkit complemented by a domain-specific taxonomy that has been validated by telecommunications experts. Although there are limitations regarding 5G and OSS coverage, the framework demonstrates promising applications for mitigation planning and incident investigation, indicating that customized threat modeling can enhance CTI alignment in legacy communication environments. Concurrently, Rajput et al. [68] investigate cyber-physical risks in Multi-Stage Flash (MSF) desalination plants. They simulate sensor and actuator attacks employing control-theoretic and finite element methods, revealing that minor disturbances can result in significant financial or structural damage. This highlights the need to incorporate CTI principles into monitoring industrial infrastructure, particularly when traditional network-based detection methods are inadequate. Addressing intelligence collection in challenging contexts, Wang et al. [69] introduce Aubrey, a Finite State Machine (FSM)-based autonomous chatbot developed to infiltrate underground IM platforms and derive actionable intelligence from fraudsters. Deployed across 150 QQ groups, Aubrey successfully collected over 7,000 artifacts—including SIM gateways and fraud toolkits—demonstrating the scalability and efficacy of proactive, conversational CTI collection in hostile environments. Lastly, Xi et al. [70] reveal a significant gap in security-targeted GNNs by presenting GTA, a dynamic graph-based backdoor attack that remains highly stealthy and transferable across datasets. With success rates exceeding 90% and minimal degradation in clean accuracy, GTA highlights the urgent need for CTI frameworks to incorporate integrity validation mechanisms for graph-based models, which are increasingly employed in malware classification and threat attribution. These contributions highlight the need to broaden CTI beyond traditional telemetry and signature-based detection. They advocate for intelligent, domain-specialized frameworks that can effectively address operational deficiencies in mobile, industrial, underground, and ML-driven environments.

In CTI, LLMs have been explored for tasks such as extracting threat actor behavior, mapping tactics to known frameworks, generating structured representations, and assisting analysts in summarizing threat reports. Despite growing interest, most LLM applications in CTI remain restricted to synthetic benchmarks or demonstrative prototypes, with limited reproducibility or integration into operational workflows. These gaps motivate the next subsection, which outlines current challenges in aligning LLM capabilities with CTI requirements.

Although LLMs derive from advancements in NLP, their adaptation to CTI introduces specific challenges: encoding threat knowledge, managing ambiguity in attacker descriptions, preserving temporal context, and aligning extracted insights with evolving intelligence schemas. Some efforts address these challenges using specialized classifiers or hybrid approaches. However, a consolidated view of how general-purpose LLMs behave in CTI-related tasks is still missing from the literature. Figure 4 presents a consolidated view of how LLMs are being directly and potentially applied in CTI workflows, while also mapping the key integration challenges—such as reproducibility, semantic ambiguity, and temporal contextualization—that hinder operational alignment.

III. METHODOLOGY: MODULAR LLM-DRIVEN SURVEY-LIKE ARCHITECTURE

The exponential growth of scientific knowledge, particularly at the dynamic intersection of cybersecurity and AI, presents several challenges in identifying gaps, trends, and disruptive research opportunities in a fast-paced and rapidly evolving context. Traditional systematic review methods, while rig-

orous, can be cumbersome and even limited in their ability to process massive volumes of literature and efficiently reveal non-obvious connections. In line with the transformative potential of LLMs in knowledge discovery highlighted in our introduction, this research investigates the application of NLP and LLMs as resources for cybersecurity, especially in potential applications at CTI activities. Also, it employs an innovative methodological approach that integrates the analytical power of LLMs into the literature review process itself.

Our methodological process was designed to transcend the limitations of large-scale manual filtering, ensuring comprehensiveness and transparency while leveraging the contextual understanding and comprehensible capabilities of LLMs to refine information selection and extraction. Instead of relying exclusively on traditional sequential filtering steps to retrieve articles directly from databases, we adopted an approach in which we first constructed a database of interest. This database serves as our initial working corpus. From this primary database, the filtering and selection of studies for in-depth analysis were facilitated and refined using algorithms based on LLMs. This hybrid process involved a more intelligent and efficient curation of the literature, optimizing the identification of articles that not only fit the thematic criteria but also contain latent research opportunities.

The methodology comprises three interrelated phases, described in the following subsections:

1) Construction of the Database of Interest: Description of the initial collection and pre-processing for the creation of the base corpus.
2) Selection and Refinement by LLMs: Explanation of how LLMs were used for filtering, ranking, and selecting the most relevant studies from the database.
3) Data Extraction and Analysis for Opportunity Discovery III-A: Details of how the information was extracted and synthesized, with an emphasis on identifying the "research opportunities" that are the "diamonds" of our work.

This methodological approach is illustrated in Figure 5. Each of these phases will be explored, providing an insight into the path taken to construct this research. Figure 6 presents the adapted PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram, visualizing the initial literature collection phases, the article filtering and selection process, and the dataset refinement conducted in this methodology [71].

A. DATA COLLECTION

The initial phase of this methodological process was dedicated to the formation of a corpus of thematically contextualized work of high scientific relevance, which would serve as input for the subsequent stages. Instead of employing traditional approaches of sequential filtering directly on scientific literature databases, our strategy prioritized the creation of a primary database of interest. Built to be intrinsically aligned

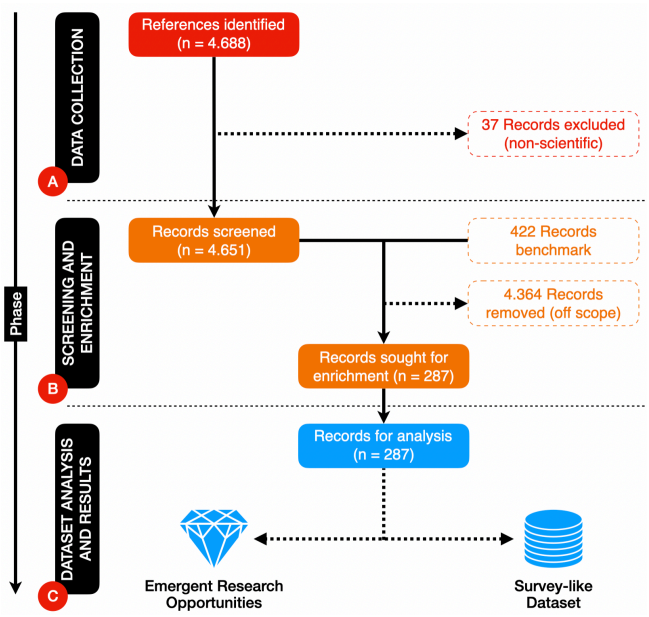

FIGURE 6. PRISMA adapted for the review methodology.

with institutional objectives, particularly the Cyber C&C and Defense Laboratory (C2-DC[2]), to which the authors of this work belong.

Initially, a substantial amount of bibliographic metadata was collected from JSON records made available by the DBLP (Digital Bibliography & Library Project). It is essential to emphasize that, at this stage, the dataset comprised exclusively conventional bibliographic information, such as titles, authors, publication locations, and publication years.

To build a contextualized corpus of high relevance to the areas of cybersecurity, CTI, AI, NLP, and LLM, the collection of papers was drawn from the following conferences and journals. It features six prominent conferences that are well-regarded in the cybersecurity community, chosen for their h5-index to guarantee scientific impact and relevance:

- USENIX Security Symposium;
- IEEE Symposium on Security and Privacy (IEEE S&P);
- ACM Conference on Computer and Communications Security (ACM CCS);
- Symposium on Network and Distributed System Security (NDSS);
- IEEE European Symposium on Security and Privacy (EuroS&P);
- ACM Asia Conference on Computer and Communications Security (Asia CCS).

In this initial collection phase, no thematic or bibliographic filter was applied. All papers published in all editions of the listed events and journals, during the period between 2018 and 2023, were considered, totaling 36 event editions. This selection aims to ensure that the corpus reflects the state of

[2]https://c2dc.ita.br/

the art and the most recent trends in the core areas of interest of the survey.

As a result of this collection process, we generated a dataset in which each row corresponds to the metadata of a paper from one of the selected conferences. This dataset initially contained 4,688 entries and was analyzed to ensure that only valid scientific publications were retained. A pre-processing step was then carried out to exclude non-scientific records (such as prefaces, discussion panels, or entries without a valid page range), stabilizing the dataset to 4.651 scientific papers that comprise our initial database.

B. SCREENING AND ENRICHMENT

The initial dataset, derived from DBLP JSON records, although comprehensive in its bibliographic coverage, was limited to bibliographic publication metadata. This granularity alone did not provide the depth to identify specific topics, methodologies, substantive contributions, or research gaps. To address this limitation and enable semantic reasoning, we implemented the Screening and Enrichment phase. This phase aimed to enrich each bibliographic work with new attributes, assisted by a systematic analysis by LLMs. This activity was developed in three complementary steps: Scope Closure and Enrichment, Fitting, and General Analysis.

B.1. Scope Closure and Enrichment: The first step in dataset improvement was crucial for enhancing the analytical value and ensuring its explicit relevance to the domain of interest, thereby preparing it for subsequent computational processing. The process began with automated abstract extraction, using XPath scraping for articles with Electronic Edition (EE) links. For PDF files, we employed structured parsers to extract the textual content, an approach particularly used for older conference publications where the availability of abstracts in a structured format may be limited. Concomitantly, a multidisciplinary relevance assessment was conducted using an LLM. This model analyzed the titles and abstracts of each article to assign binary labels indicating their relevance to specific research subdomains (Cybersecurity, CTI, AI, NLP, and LLM). Based on the labels, the papers were screened, retaining only those publications that demonstrated alignment with the research interests established in the survey (CTI, NLP, and LLM).

B.2. Fitting: With the dataset now filtered and enriched with abstracts, the second stage of the phase focused on extracting high-level information for analysis and scrutiny according to the criteria of this work. To this end, a Question-and-Answer analysis was employed. This process aimed to extract characteristic information about scientific aspects (research objectives, methodologies employed, contributions, limitations identified by the authors, and suggested future directions). Additionally, a refinement phase reprocessed the generated responses to improve clarity and analytical value, thereby reducing the subjectivity inherent in the raw reactions of generative models and synthesizing insights into more computable and enumerable concepts.

B.3. General Analysis: This step organizes the dataset and uncovers thematic patterns to support further analysis. The General Analysis process begins with LLM, which extracts points, counterpoints, and gaps. Then, a relevance assessment measures the degree of alignment of each work with a specific target theme, resulting in a categorical relevance score (High, Medium, Low, None) accompanied by a corresponding textual justification. Subsequently, a category labeling process aims to group works into categories, developing thematic clustering.

C. DATASET ANALYSIS AND RESULTS

The final phase of the analytical workflow focuses on extracting and synthesizing research opportunities from the dataset. This process aims to identify future research propositions while also refining and consolidating them, presenting a set of high-impact directions. Section IV presents in detail the grouping, analysis, and discussions of the findings from the dataset. The result of this phase is a strategic set of emerging research directions, synthesized from multiple sources and deemed impactful in the investigated domain.

D. OVERVIEW

Based on these insights, the Relevance Filtering and Scope Narrowing step concentrated on papers that demonstrated strong and consistent relevance to the most clearly defined subdomains. We specifically retained only those entries marked as relevant (Y) to CTI, NLP, or LLM, as specified in the expression 1.

$$(LLM = 1) \vee (CTI = 1) \vee (NLP = 1) \tag{1}$$

An analysis of the dataset, as delineated in the Equation 1, reveals patterns of thematic distribution and intersection across the selected research subdomains. Notably, a subset of documents demonstrated multi-label relevance, consisting of 9 papers that fulfilled the criteria $(CTI = 1) \wedge (NLP = 1)$, one entry that satisfied $(CTI = 1) \wedge (LLM = 1)$, and 28 instances that adhered to $(NLP = 1) \wedge (LLM = 1)$. This suggests a considerable degree of conceptual overlap within the corpus. Conversely, another collection of publications exhibited exclusive relevance to a single domain: specifically, 44 papers were uniquely associated with $(CTI = 1) \wedge (NLP = 0) \wedge (LLM = 0)$, 84 with $(NLP = 1) \wedge (CTI = 0) \wedge (LLM = 0)$, and 121 with $(LLM = 1) \wedge (CTI = 0) \wedge (NLP = 0)$.

These observations reveal a pronounced thematic focus in the areas of NLP and LLM, whereas the field of CTI demonstrates a narrower presence and connections. These distribution trends not only define the current research landscape but also imply the development of epistemological ties between language technologies and cybersecurity intelligence.

Figure 7 displays the yearly distribution of publications in the CTI, NLP, and LLM subdomains from 2018 to 2023. Starting in 2021, there is a rise in the volume of papers for the LLM subdomain, along with a smaller increase in NLP. CTI-related publications remained steady during this

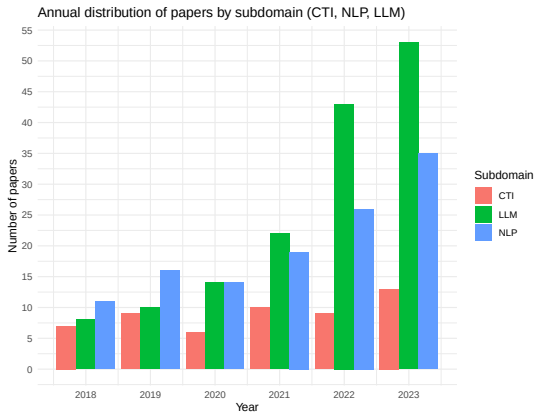

FIGURE 7. Annual distribution of papers by subdomain.

time, with a slight uptick in 2023. These results suggest a recent surge in research involving LLM models and reinforce the anticipation that the proposed methodology will uncover, among other aspects, emerging research opportunities at the convergence of LLM and CTI. Strategic research opportunities were identified through a semantic analysis of the 287 vetted scientific papers, which included thematic summaries, main contributions, critical counterpoints, and related research gaps. Evaluating these annotated aspects through AI-assisted stages enabled tracing the technical core of each study and its hidden potential.

Throughout this refinement, papers with overlapping gaps, core ideas, or complementary themes were grouped and synthesized to create a cohesive framework. This gradual aggregation uncovered convergence patterns across domains, which were not always clearly stated in the original texts. These intersections, referred to as opportunity spaces, were analysed, targeted, and principal opportunities identified. The outcome of this process was a vertical consolidation pipeline that transformed an initial collection of 287 annotated opportunities into a refined selection of 8 diamonds. These "diamonds" signify formulations that capture a wider scientific potential, integrating relevance, feasibility, and innovation, arising from the convergence of themes rather than remaining separated within them.

E. RELEVANCE CLASSIFICATION AND HUMAN VERIFICATION

The classification stage was designed to replicate, while also accelerating, the traditional survey methodology, where researchers manually assess the relevance of papers based on their titles, abstracts, and keywords. In our case, the binary classification (YES/NO) was performed by a general-purpose LLM via API, but still grounded in that same semantic context, especially concerning the domain of CTI and adjacent areas like NLP and AI.

To guide the model, we provided a reference set of 422 manually labeled papers. These served as a benchmark for prompt design and also as a comparative ground truth. The model's classifications were further audited through sampling—both randomly and by checking papers known to be relevant—over multiple iterations. Each cycle led to refined prompt strategies, which gradually improved the reliability of the strategy.

We also evaluated the classification performance quantitatively by comparing model outputs to human annotations. The resulting accuracy rates were encouraging: the approach achieved an overall accuracy of 94.22% when benchmarked against manual labels. This included high alignment in structured subdomains, such as NLP and LLM; however, performance in CTI was more uneven, possibly due to fuzzier boundaries and less distinctive language patterns.

While this does not eliminate the need for human review, it does offer a scalable method for initial triage, especially useful when facing thousands of publications. The resulting dataset, derived from this process, is openly available for inspection and reuse at IEEE Dataport; see [71].

IV. DISCUSSION OF STRATEGIC RESEARCH OPPORTUNITIES

This section outlines the main strategic and emergent research directions identified in our survey. Each subsection addresses a topic corresponding to a "diamond," highlighting its context and specific considerations within CTI.

A. AI-DRIVEN CYBERSECURITY RISK ASSESSMENT

In the evolving cybersecurity landscape, CTI has emerged as a critical discipline focused on collecting, analyzing, and operationalizing information about threats to anticipate, detect, and respond to malicious activities. With the exponential growth of threat data and the increasing sophistication of attacks, traditional CTI methods often fall short in extracting actionable insights promptly. In this context, NLP and LLMs have become transformative tools for automating the extraction, interpretation, and correlation of threat intelligence from vast and diverse data sources, including incident reports, vulnerability disclosures, threat feeds, and malware analyses. By enabling machines to understand the semantics and context of unstructured textual information, NLP and LLMs enhance CTI workflows with entity recognition, relationship mapping, and narrative reconstruction, thereby supporting faster decision-making and more resilient cyber defense strategies.

Works such as the Fereidooni et al. [72] and De La Torre Parra et al. [73] emphasize collaborative threat intelligence without compromising data sovereignty. Fereidooni et al. demonstrate, using FedCRI, the viability of Federated Learning (FL) in sharing cyber-risk patterns among mobile service providers, thereby mitigating the privacy limitations of traditional CTI that rely on static indicator exchanges. Using GRU-based models updated through FL, FedCRI preserves user-level confidentiality and enables dynamic risk response strategies such as contextual authentication and feature restriction. The system's scalability (tested on data from 23.8 million users) underscores its potential to support real-time CTI workflows with global reach. Similarly, De La

TABLE 2. Analysis of AI-Driven Solutions in Cybersecurity Risk Assessment

System / Author(s) - Focus	Type of AI/NLP	Data Source (Interpretation)	CTI lifecycle	Key Differential	Limitations/Challenges
[72] (FedCRI) - Risk sharing without exposing private data	FL + GRU	Mobile telemetry (Low)	Detection & Sharing	Privacy-preserving, scalable	Latency, synchronization
[73] - Forensic interpretation of distributed logs	FL + Transformer with attention	Syslogs (High)	Forensics & Indicator validation	Attention-based interpretability	Computational cost
[74] (Forecast) - Extraction / prioritization of malware capabilities	Symbolic Execution + empirical data	Memory snapshots (Medium)	Threat indicator enrichment	"Degree of concreteness" prioritization	Scalability to large datasets
[75] (ATLAS) - Reconstruction of attack narratives	NLP + Deep Learning + Causality analysis	Audit logs (High)	Timeline enrichment, lateral movement	Temporal, semantically enriched narratives	Generalization to other domains
[76] (DISTDET) - Large-scale APT detection	Distributed anomaly detection + Semantic correlation graphs	Host telemetry (Medium)	Real-time distributed detection	Centralized summary graphs, false positive reduction	Multi-organizational integration
[77] (HOLMES) - Mapping to MITRE ATT&CK	Provenance Graph + NLP	Information flows (High)	Automated scenario reconstruction, kill-chain	Detailed semantic mapping of attack steps	Log noise sensitivity
[76] - Adversarial robustness of AI models	LLM + Adversarial Attacks	Training data for LLMs (-)	Secure AI development	Highlights new attack surfaces on LLMs	Lack of systematic defenses

Torre Parra et al. use the Interpretable Federated Transformer Log Learning to advance this paradigm by pairing FL with transformer-based architectures for syslog analysis. What sets it apart is the inclusion of a token-level attention-based interpretability module, allowing forensic analysts to attribute model decisions to specific log features. This interpretability is essential for CTI applications where transparent decision-making is needed for indicator validation and incident reconstruction.

Alrawi et al. [74] focus on interpreting behavior and recovering threat narratives. Via the Forecast System, Alrawi et al. employ symbolic execution with real execution data to extract malware capabilities from forensic memory snapshots. Its "degree of concreteness" metric offers a novel prioritization mechanism, making symbolic analysis tractable even for obfuscated malware. Although not explicitly framed as a CTI system, Forecast provides high-fidelity outputs that can enrich CTI databases with capability-level indicators, thereby enhancing both tactical response and long-term threat profiling. Alsaheel et al. [75] with the ATLAS, on the other hand, reconstruct attack stories from audit logs using a hybrid of NLP, DL, and causality analysis. ATLAS bridges raw telemetry and human-readable attack narratives by correlating temporal log sequences and extracting semantically enriched patterns. Its architecture supports targeted investigation (starting from known symptoms) and iterative threat expansion—core functions in CTI workflows such as timeline enrichment and lateral movement analysis.

Dong et al. [78], Milajerdi et al. [77], and Schuster et al. [76] contribute to the advancement of CTI by exploring high-fidelity detection and the exposure of emergent threat surfaces at the system and model levels. Dong et al. introduce DISTDET, a distributed system for detecting Advanced Persistent Threats (APTs) that combines lightweight local anomaly detection with centralized semantic correlation using Alarm Summary Graphs. By integrating host-specific models with a scalable ranking mechanism, DISTDET effectively reduces false positives while maintaining precision. It applies to enterprise CTI settings where large-scale telemetry, contextual analysis, and real-time responsiveness are essential. In a complementary vein, Milajerdi et al. present HOLMES, a system that maps suspicious information flows from audit logs to MITRE ATT&CK Tactics, Techniques, and Procedures (TTPs) through compact provenance graphs. The resulting High-level Scenario Graphs and threat tuples enable structured semantic interpretation of attack sequences, providing tactical insight and supporting integration into CTI processes such as automated scenario reconstruction and kill-chain analysis. Interpretability is a non-negotiable requirement for LLM deployment in high-stakes CTI environments. State-of-the-art techniques such as SHAP, LIME, and attention-based visualization [56], [79], [80] have been leveraged to increase transparency, trust, and auditability in LLM outputs. These approaches enable practitioners to attribute model decisions to concrete input features, supporting both regulatory compliance and operational confidence.

Schuster et al., in contrast, expose a novel and increasingly relevant concern regarding CTIs: the vulnerability of LLMs to poisoning attacks. By manipulating training data or fine-tuning lifecycles, adversaries can bias code completion models toward insecure defaults, introducing risks in AI-assisted software development environments. Although not detection tools in the classical sense, these findings underscore the adversarial potential inherent to LLM-based systems and the necessity for CTI-aware defense mechanisms in AI development workflows. Together, these works highlight the importance of provenance modeling, semantic correlation, and adversarial robustness in developing AI-driven CTI systems that are responsive and resilient to emerging attack vectors. To facilitate comparison between the leading recent solutions, Table 2 presents a multidimensional analysis of AI methods applied to risk assessment in CTI, highlighting differences in approach, integration, and limitations.

B. SECURE IOT ECOSYSTEMS USING AI

The rise of IoT devices has led to improvements in automation and connectivity, but it has also introduced numerous security risks. These devices typically operate in decentralized, resource-constrained environments, making conventional security solutions inadequate. To address these issues, research explores the integration of AI techniques to design advanced intrusion detection and prevention systems tailored for IoT. By training AI models on real-time IoT telemetry, the goal is to detect anomalies, unauthorized access attempts, and coordinated attacks, thereby enhancing system resilience and data confidentiality.

Studies have explored complementary techniques to support this objective. Feng et al. [81], for instance, developed ARE—a rule-based discovery engine that identifies and classifies IoT devices using application-layer responses and rule mining without requiring labeled training data. While not an intrusion detection system per se, ARE enables scalable IoT device fingerprinting, which is fundamental to asset visibility, vulnerability mapping, and exposure assessment—essential tasks in proactive security architectures. In another vein, Cong et al. [82] introduced SSLGuard, a watermarking system designed to protect self-supervised learning (SSL) encoders. Though not directly applied to IoT security, the method's robustness against model stealing and tampering has implications for protecting AI-based detection models embedded in edge devices.

The work of Ryan et al. [83] does not involve behavioral anomaly detection, as previously demonstrated, but rather exposes a cryptographic vulnerability in RSA-based SSH key generation. The authors demonstrate that passive adversaries can extract private keys from faulty signatures using lattice-based attacks, revealing systemic weaknesses across IoT and networking equipment vendors. These findings are particularly important for CTI, as they enable the monitoring of long-lived cryptographic assets. Similarly, Hussain et al. [84] do not propose a sensor-actuator interaction model, but COINN, a privacy-preserving crypto/ML framework that enables secure DNN inference. The system introduces optimizations, including quantized secure computation and automated configuration tuning. Although the primary contribution is in secure AI computation, its techniques could be repurposed for privacy-sensitive threat detection in distributed IoT settings.

Regarding firmware security, Han et al. [85] presented UNICORN, a host-based intrusion detection system that leverages system provenance to identify APTs in real-time. Its graph sketching and behavioral modeling techniques address long-term, stealthy attacks—a significant advance over traditional short-range anomaly detectors. These contributions are especially relevant for endpoint-level IoT monitoring. Chen et al. [86] developed Atomic, a framework for detecting vulnerabilities in LTE NAS protocols through the semantic analysis of documentation. By automating test case generation from technical standards, Atomic uncovers exploitable flaws in communication logic. Although focused on mobile networks, its methodology (NLP-driven FSM modeling) is transferable to IoT firmware and protocol analysis.

Zhao et al. [87] introduce UVSCAN, a binary-level framework that utilizes NLP and logic-based static analysis to detect misuse of Third-Party Components (TPCs) in IoT firmware. Rather than performing Distributed Denial of Service (DDoS) detection on Application Programming Interfaces (APIs), UVSCAN aligns API documentation with binary code using RoBERTa-based MRC and logic inference, enabling detection of misuses that violate API intent—vital for supply-chain risk analysis and CTI. Chang et al. [88] do not address bootstrapping with Physical Unclonable Function (PUFs). Instead, they propose DSCORR, a DNS session correlation system for user re-identification via Word2Vec-based embeddings and auto-thresholding. To defend against DNS-based tracking, they introduce LDPRESOLVE, a privacy-preserving DNS resolver. These findings contribute to the anonymization and privacy-preserving telemetry space, which is a growing concern in large-scale IoT surveillance environments.

Liu et al. [89] introduce a neural Trojan horse attack for pre-trained models, demonstrating how adversaries can inject behaviors into models without access to training data. This has severe implications for the integrity of AI-based security models deployed in IoT systems, especially in FL or edge inference. Rimmer et al. [90] do not focus specifically on fingerprinting smart devices. Instead, they study website fingerprinting over Tor using DNNs to automate traffic classification. While this work is primarily focused on privacy and anonymity in encrypted networks, the methods can also inform traffic analysis and behavioral profiling in encrypted IoT environments. To provide a comparative overview of the key AI-based strategies applied to secure IoT ecosystems, Table 3 summarizes the main contributions, techniques, and datasets used in the selected studies discussed in this subsection.

1) Knowledge Extraction from Networking Log Files: a case study

Among the wide spectrum of CTI data sources, network log files represent one of the most technically challenging and operationally valuable modalities for automated analysis. These logs—generated continuously by routers, gateways, and intrusion detection systems—encode essential information about packet flow, timestamps, signal parameters, and device identifiers. Their structure, however, is typically unstandardized and massive in scale, rendering manual analysis infeasible and traditional parsing methods inefficient.

Recent work by Siino, Giuliano, and Tinnirello [91] demonstrates a practical approach to knowledge extraction from networking logs using LLMs. In their study, the authors leveraged the few-shot learning and Chain-of-Code (CoC) capabilities of LLMs—prompting the model either to directly interpret logs in natural language or to generate executable Python code to automate the analysis. Both techniques were tested on LoRa network logs, which include metrics such as RSSI, SNR, bandwidth, spreading factor, and device EUI. Their results reveal that while direct prompting produces

TABLE 3. CTI-Enhanced AI Mitigations for Key IoT Vulnerabilities

Vulnerability	Reference and Description	CTI-Based Mitigation
Device Visibility	ARE enables device fingerprinting via rule mining [81]	CTI augments inventory and exposure mapping
Model Integrity	SSLGuard resists model theft [82]; Trojan injection in DNNs [89]	CTI tracks AI threat indicators and injection signatures
Crypto Weakness	RSA SSH key flaw via lattice attacks [83]	CTI supports long-term crypto asset monitoring
Firmware Exploits	Semantic bugs in LTE NAS (Atomic) [86]; TPC misuse via UVSCAN [87]	CTI enables supply chain risk detection
Privacy Leakage	DNS-based re-ID with DSCORR [88]	CTI-informed traffic anonymization strategies
APT Detection	UNICORN uses system provenance for APT detection [85]	CTI facilitates correlation and behavior baselining
Traffic Profiling	DNN-based encrypted traffic classification [90]	CTI supports pattern-aware traffic fingerprinting
FL Inference Leakage	COINN ensures secure DNN inference [84]	CTI guides privacy-aware AI configuration

acceptable syntactic accuracy, the CoC approach achieves perfect semantic precision (Precision = 1.0, Recall = 1.0) by allowing the model to reason through code execution.

This methodology highlights an emerging paradigm for CTI automation: combining LLM-driven semantic understanding with code synthesis to achieve interpretable, verifiable outcomes. When applied to large-scale cybersecurity infrastructures, such approaches could transform Security Information and Event Management (SIEM) systems, enabling them to extract actionable intelligence from unstructured logs in near real-time.

Integrating LLMs in this context introduces new research directions related to real-time inference, model fine-tuning for network telemetry, and trustworthy integration with existing operational systems. These challenges illustrate one of the most promising ''diamonds'' for future CTI research—bridging the gap between unstructured data overload and structured, machine-verifiable intelligence.

Beyond data volume, the capacity to interpret and correlate heterogeneous sources defines the next frontier of intelligent threat analysis. The transition from reactive data collection to proactive pattern discovery hinges on integrating LLMs into dynamic pipelines capable of real-time adaptation.

C. APPLYING AI IN IDS

The dynamic and evolving nature of cyber threats demands creative approaches to Intrusion Detection System (IDS). Traditional rule-based IDS often fails to identify novel or sophisticated attacks, such as zero-day exploits or stealthy adversarial actions. In response, researchers have increasingly focused on integrating AI, in particular ML, DL, and LLMs, into IDS architectures. This integration enhances anomaly detection capabilities and reduces false positives, thereby improving operational effectiveness and enabling real-time adaptation to emergent threats.

AI-based IDS aims to overcome the limitations of static signature detection by employing adaptive models that can generalize from known threats to detect previously unseen attack vectors. For example, frameworks such as NODOZE [92] employ anomaly score propagation and graph-based behavioral partitioning to prioritize alerts and reduce analyst fatigue. By learning from enterprise-scale telemetry and applying network diffusion algorithms, such systems triage alerts with minimal human supervision—an ideal capability within CTI lifecycles where threat volumes are unmanageable without automation.

ML algorithms also face adversarial challenges. SAGE [93], a self-attention distillation mechanism for DNNs, demonstrates how internal representations of models can be purified to eliminate hidden backdoors without degrading prediction accuracy. Such approaches can be adapted in CTI to sanitize LLMs and transformer-based classifiers that have been compromised through data poisoning or adversarial pretraining, thereby ensuring the integrity of threat detection modules embedded in critical infrastructure. Similarly, D-DAE [94] introduces mechanisms to reconstruct disrupted outputs from black-box models using generative approaches, enhancing resilience against anti-extraction defenses. These recovery mechanisms hold promise for CTI, where security researchers often interact with opaque or proprietary systems and require reliable methods for semantic reconstruction of threat data, especially when investigating evasive adversaries or compromised systems.

AI also contributes to understanding and detecting sophisticated threats at the binary and behavioral levels. The PELICAN framework [95] illustrates how transformer models can be manipulated through instruction-level backdoor triggers while maintaining operational semantics. Although initially designed for binary code analysis, this technique can be leveraged in CTI to assess model vulnerabilities and enhance the robustness testing of malware classifiers, especially when attackers exploit compiler-specific syntax regularities. NLP and LLM-based techniques expand the scope of IDS beyond raw traffic or binary analysis to higher-level cognitive threat behaviors. In the realm of password security, PassBERT [96] leverages BERT-style transformers trained on leaked credentials to outperform existing models in conditional and targeted password guessing. Such models can be integrated into CTI frameworks to predict attacker credential reuse strategies or to augment behavioral profiling of adversaries based on leaked datasets.

Meanwhile, Cidon et al. [97] provide a practical case of hybrid AI for detecting business email compromise attacks by fusing metadata-based impersonation detection with content-

based classifiers. While not an IDS per se, its architecture exemplifies a supervised learning approach tightly coupled with real-world operational constraints—an essential attribute for CTI systems managing phishing intelligence or adversary infrastructure monitoring. Beyond detection, AI is also being used to infer model behavior and expose privacy vulnerabilities. For instance, the encoder-decoder-based analysis presented in [98] demonstrates how output shifts in online learning models can be exploited to reconstruct private data. This introduces both a threat and an opportunity: on one hand, it necessitates better defenses in CTI tools; on the other, it could enable analysts to infer attacker updates in real-time from observable system behaviors.

Fundamentally, transformer models are being increasingly used not only for text but also for model extraction across various domains. Battis and Payer [99] reveal how vision transformers can effectively replicate proprietary CNNs through knowledge distillation, bypassing common defenses. These findings can be instrumental in CTI environments for reconstructing unknown classifiers deployed by threat actors, allowing defenders to simulate adversarial use cases preemptively. The detection of similarity across code variants is another area where AI significantly enhances IDS capabilities. GESS [100], for instance, utilizes graph embeddings for the scalable detection of vulnerable code across heterogeneous binaries, achieving up to a 60% improvement in recall over prior methods. When embedded within CTI processes, such tools allow analysts to detect code reuse across malware families and identify variant propagation through threat ecosystems.

At last, techniques such as ImU [101], although primarily developed for physical impersonation attacks, highlight the growing convergence between adversarial ML and biometric security. In CTI applications, similar generative embedding techniques can be used to evaluate the impersonation resilience of biometric authentication mechanisms or to train threat detection models that are robust against adversarial face spoofing.

1) The Continued Relevance of Text Preprocessing for LLM-Driven CTI

Although modern Transformers have reduced the dependence on handcrafted features, treating text preprocessing as optional is both premature and misleading. In CTI, where inputs are noisy and heterogeneous—ranging from incident reports to dark-web chatter—preprocessing remains decisive for semantic coherence and analytical accuracy. Siino, Tinnirello, and La Cascia [102] show that the choice and ordering of techniques such as stop-word removal, lemmatization, stemming, spelling correction, negation handling, and POS-aware sequencing can change classification accuracy by up to 25%, even for state-of-the-art Transformers (e.g., XLNet). They also report cases where a simple Naïve Bayes outperforms a Transformer when backed by an appropriate preprocessing pipeline. In line with these findings, our CTI architecture explicitly includes a preprocessing layer responsible for cleaning, normalizing, and enriching raw text prior to LLM ingestion, thereby improving robustness, efficiency, and reproducibility in downstream intelligence workflows.

D. ENHANCED CTI THROUGH NLP

Harnessing NLP to enhance CTI represents a promising and timely avenue. This approach advocates for the development of NLP-driven systems that can effectively extract, analyze, and prioritize threat intelligence from unstructured and informal text sources, including dark web forums, social media, and security blogs. Such as semantic comprehension, NER, emotion identification, and dependency analysis enhance both the precision and contextual relevance of threat detection and behavioral analysis. To illustrate how NLP techniques can be integrated into CTI workflows, Figure 8 presents a conceptual workflow that highlights the stages of data ingestion, processing, and the generation of actionable intelligence.

Recent research supports the introduction of SteinerLog, a system for reconstructing APTs using provenance graphs and CTI-driven IoA rules, as presented by Bhattarai and Huang [103]. Although not originally intended for NLP processing of open text, its use of structured threat representations based on MITRE ATT&CK and integration of public threat intelligence (e.g., Sigma rules) exemplifies how CTI lifecycles can benefit from rich semantic context and structured correlation of behavioral evidence. Zhu and Dumitras [104], in contrast, focus directly on unstructured text processing. Their system, ChainSmith, employs dependency-based word embeddings and NER to extract Indicators of Compromise (IOCs) from technical reports and classify them into stages of malware campaigns. This work provides a foundation for integrating NLP with CTI lifecycles by linking linguistic features to telemetry data, enabling the profiling of large-scale campaigns and the semantic annotation of threat indicators.

Expanding the interpretability of CTI systems, Shin et al. [105] proposed W2E, which detects emerging cybersecurity events on Twitter by tracking semantically significant word-level shifts rather than tweet volume. Through POS tagging and clustering over new terms, their method identifies emergent threats early and with high precision, highlighting the potential of NLP in real-time CTI alerting. Kashapov et al. [106] also emphasize interpretability, introducing a transformer-based lifecycle for aiding users in phishing email detection. By integrating extractive summarization, emotional trigger identification, and intent detection, their system generates semantically focused representations that expose psychological manipulation in phishing content. While primarily user-facing, these techniques can support CTI workflows that require human validation or cognitive profiling of threat narratives.

On the structural side, Nadeem et al. [107] introduced SAGE, a visual analytics system that models attacker behavior from raw alerts using probabilistic automata. Although it does not rely on NLP per se, SAGE provides structured semantic abstractions of adversarial behavior, which are key

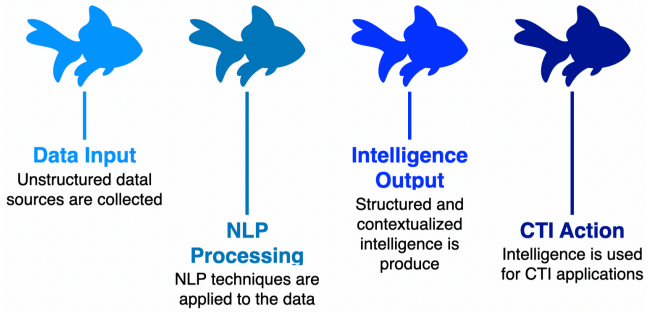

FIGURE 8. **Workflow for Improving CTI with NLP.**

to integrating NLP-derived threat narratives with behavioral graphs, thereby enabling hybrid CTI lifecycles. Zeng et al. [108] developed SHADEWATCHER, a graph-based detection system leveraging contextual embeddings and system audit logs to assess the adversarial likelihood of interactions. While not based on open text, its semantic modeling of entity behavior and integration of analyst feedback mirror CTI objectives in behavioral profiling and anomaly detection, demonstrating how graph representations can complement NLP in CTI.

Hendler et al. [109] examined NLP and DL techniques for detecting obfuscated PowerShell commands. Their results highlight that raw text embeddings, particularly CNN-based models, can enhance the detection of adversarial scripts—an application directly relevant to CTI systems that monitor command-line abuse. Milajerdi et al. [110] extended this line of research with Poirot, a system that aligns CTI-derived query graphs with provenance data through inexact graph pattern matching. Poirot connects unstructured threat descriptions with concrete system behavior by modeling CTI indicators from text and aligning them with audit records, demonstrating a mature integration of NLP-informed intelligence into threat hunting lifecycles.

In another approach, Ryan et al. [111] presented MalONT2.0, a framework that encodes Android malware concepts into RDF triples based on unstructured reports, blogs, and tweets. This work directly supports CTI goals by transforming linguistic threat narratives into structured, machine-actionable formats and integrating them into graph-based knowledge systems for forensic and strategic analysis. Heijden and Allodi [112], although not specifically focused on NLP, modeled cognitive vulnerability triggers in phishing emails through supervised topic modeling. Their behavioral approach, grounded in psychological principles, offers a unique layer of analysis that can be enriched by NLP-based emotion or intent detection tools in CTI systems.

These studies reinforce the research proposition that NLP can enable the extraction and use of CTI from diverse, noisy, and informal sources. They show that automating, scaling, and contextualizing CTI generation is feasible by combining semantic processing, structured representation, and behavior modeling. However, challenges persist, including linguistic variability, multilingualism, informal text structure, and computational demands for real-time inference, especially with transformer models like BERT or GPT in latency-sensitive environments. This proposal diverges from standard CTI practices by advocating for the integration of sentiment analysis and semantic relationship modeling, enabling more context-sensitive, cognitively enriched threat evaluations. It enables CTI systems to function effectively where signature-based methods fail, leveraging the flexibility and abstraction capabilities of modern NLP.

E. USING LLMS FOR CTI

The cyber threat landscape presents a significant challenge to CTI operations, especially due to the overwhelming volume and complexity of unstructured data. Traditional CTI approaches—relying heavily on manual analysis or semi-automated tools—struggle to process this deluge of threat reports, logs, and telemetry at scale. Advances with NLP and the emergence of LLMs offer a transformative opportunity: to automate the extraction of actionable intelligence from massive textual corpora, thereby accelerating threat detection, enhancing precision, and mitigating information overload. To operationalize the integration of LLMs into CTI workflows, Figure 9 depicts the modular architecture of an LLM-assisted pipeline, illustrating its key components from data collection to opportunity discovery.

LLMs are well-suited for CTI due to their ability to capture contextual semantics, model dependencies in long documents, and generalize across diverse text formats. Their application has begun to show promise in various CTI tasks, ranging from log-level analysis to behavior extraction, training data reconstruction, and protocol understanding. The EXTRACTOR framework [113] exemplifies the capabilities of LLM-assisted NLP for parsing unstructured CTI reports into semantically rich behavior graphs. By combining coreference resolution, semantic role labeling, and summarization, EXTRACTOR constructs provenance graphs capable of automated threat matching—a significant leap in structuring verbose threat narratives into machine-readable intelligence.

Similarly, AIRTAG [114] utilizes a customized BERT tokenizer to classify raw logs without requiring manual labeling or graph construction. By embedding log entries as semantic units and applying one-class SVM classifiers, AIRTAG surpasses traditional forensic systems in speed and accuracy, confirming that fine-grained NLP embeddings are viable for precise and scalable attack detection. Another notable development is the ProvG-Searcher system [115], which enables threat behavior search by encoding provenance graphs using graph embeddings. It allows CTI analysts to query using natural language descriptions translated into graph-based structures, showcasing the synergy between LLMs, graph learning, and CTI for hypothesis-driven threat hunting.

Moreover, NLP models are now being used not just for detection but for the synthesis and simulation of threat behaviors. The work by Charmet et al. [116] proposes an adversarial URL generation pipeline that utilizes GANs to simu-

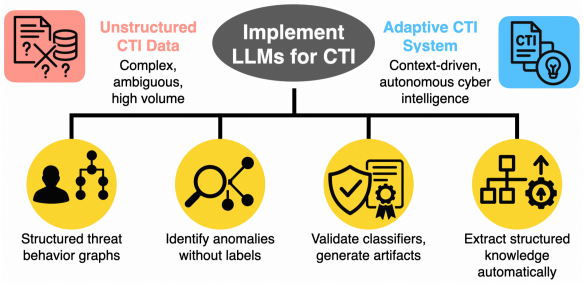

FIGURE 9. Modular architecture of the LLM-assisted pipeline.

late polymorphic phishing links, which is helpful in training LLM-based classifiers and testing detection systems under adversarial conditions.

The potential of LLMs in CTI also raises novel security risks. Lin et al. [117] warn of model extraction attacks (e.g., QUDA) that can clone DL models via generative adversarial and reinforcement learning techniques, even under strict query limits. Such threats necessitate reconsideration of model exposure policies in CTI lifecycles. From a data integrity perspective, LLM-driven CTI workflows must also account for data poisoning and model compromise. Shan et al. [118] present a framework that retrospectively identifies poisoned training samples using gradient-based clustering and unlearning, enabling evidence-based remediation for compromised threat detection models.

Beyond immediate detection, LLMs facilitate the recovery of structured security knowledge from unstructured sources. Pacheco et al. [119] demonstrate this in protocol analysis, using BERT embeddings and neural sequence labeling to extract FSMs from technical RFCs. This capability directly supports automated test generation and adversarial scenario simulation—key activities in proactive CTI. The use of pretrained language embeddings has also been effective in detecting evasive threats. Hendler et al. [120] demonstrate that deep contextual embeddings trained on unlabeled PowerShell scripts can accurately identify obfuscated and fileless malware, outperforming traditional detection systems in operational environments.

Lastly, the Advance framework [121] demonstrates how semantic assumptions in API documentation can be automatically extracted and utilized to synthesize verification code, thereby enabling the automated detection of misuse in real-world software. This demonstrates a practical pathway for bridging documentation with executable CTI analysis workflows.

Taken together, these studies affirm that the integration of LLMs and NLP into CTI offers substantial benefits: scalable threat detection, enriched data structuring, adversarial simulation, and automation of analytical workflows. They also expose critical challenges in privacy, model security, and data integrity, which must be addressed through privacy-preserving training methods (e.g., differential privacy) and robust auditing frameworks [122].

F. NLP FOR SOCIAL ENGINEERING DETECTION

Social engineering attacks exploit human cognitive and emotional vulnerabilities rather than technical flaws, making them difficult to detect with conventional cybersecurity mechanisms. In response, NLP and LLMs offer promising avenues for detecting these manipulative behaviors by analyzing linguistic cues and psychological patterns embedded in text or speech.

While traditional CTI platforms focus on malware signatures, IP indicators, and technical exploits, they often overlook the subtle linguistic manipulations that characterize social engineering attacks. By leveraging NLP, these platforms can be enhanced to analyze discourse features, including sentiment, politeness strategies, rhetorical structures, and deception markers. Masked Language Models (MLMs) such as RoBERTa have been shown to effectively detect adversarial text perturbations by identifying deviations from the language manifold [123]. This detection capacity could be repurposed in CTI systems to flag unnatural or psychologically coercive language patterns indicative of phishing or pretexting. Moreover, LLMs are increasingly deployed in CTI for tasks such as threat summarization, incident report synthesis, and analyst support. However, their open-ended generative nature exposes them to novel threats. The ability to reverse-engineer decoding strategies in LLM APIs, as demonstrated by [124], reveals how adversaries could infer defensive configurations or reconstruct threat response strategies embedded in CTI tools. Similarly, backdoor injection techniques such as LISM [125] and model spinning [126] expose CTI systems to the risk of stealthy manipulations, where LMs can be covertly altered to exhibit biased or deceptive outputs triggered by stylistic or semantic inputs.

These risks are not merely theoretical. Tools like ToxicBuddy [127] demonstrate that non-toxic queries can be crafted to elicit toxic outputs from chatbot systems, potentially undermining the credibility of CTI conversational agents. Similarly, the LogNormMix-Net framework [128]—designed to simulate realistic messaging environments—can be adapted to generate synthetic social engineering scenarios for red-teaming and training purposes. Such simulated dialogues could help develop robust detectors that distinguish genuine from malicious intent in organizational communication streams.

The threat is compounded by adversarial strategies, such as homograph substitution and dynamic sentence triggers, which remain invisible to both users and automated systems [129]. These methods can embed hidden triggers in CTI data lifecycles, subtly altering classification outcomes or inducing bias in summarization tasks. Even more alarming, [130] shows that backdoors can be introduced during pretraining without access to downstream labels, implying that even widely used public models may harbor latent manipulations that are exploitable in CTI applications.

NLP models also present a dual-use risk: while they enhance detection, they can also inadvertently leak sensitive information. For instance, [131] demonstrates that embed-

dings generated by general-purpose LLMs can reveal identifiable patterns or keywords, thereby compromising the confidentiality of threat reports or entity recognition tasks in CTI platforms. Similarly, the LipFuzzer framework [132], though initially designed for voice assistants, highlights how semantic fuzzing can be used to test and harden the intent recognition layer of CTI interfaces against manipulation.

Despite these challenges, the convergence of NLP and CTI offers a transformative opportunity to address the long-standing "human factor" in cybersecurity. By focusing on language as both a vector and a signal of attack, researchers can develop LLM-powered detectors trained on social engineering corpora to recognize exploitative discourse structures. Such models must incorporate behavioral context, pragmatic inference, and conversational dynamics to distinguish benign anomalies from intentional manipulation.

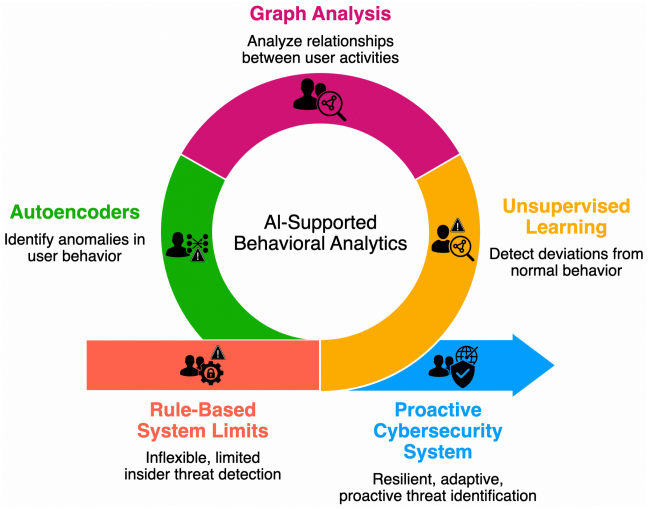

FIGURE 10. AI-Driven Insider Threat Detection

The integration of NLP and LLMs into CTI frameworks represents a significant advancement in defending against social engineering. By leveraging manifold-sensitive models, detecting linguistic anomalies, and anticipating adversarial misuse of language, researchers can equip CTI platforms with the capacity to recognize manipulative intent at scale. However, these advancements must be accompanied by rigorous auditing, adversarial testing, and privacy-preserving mechanisms to safeguard against model exploitation. Ultimately, addressing the human dimension of cyber threats requires a linguistic lens—one that combines the technical precision of NLP with the behavioral insights central to social engineering detection.

G. BEHAVIORAL ANALYSIS FOR INSIDER THREAT DETECTION

Insider threats continue to pose a complex challenge in cybersecurity, often originating from authorized users whose malicious actions are disguised within legitimate activities. This strategy combines AI methods with behavioral analysis to identify unusual patterns that indicate potential insider threats. By utilizing ML models to examine access logs, traces of user behavior, and system activities, it seeks to address the rigid constraints of conventional rule-based detection systems. It offers a more adaptive and context-sensitive approach that can uncover nuanced signs of internal misuse. Insider threats remain a complex and persistent challenge in cybersecurity, often involving malicious actions performed by authorized users and concealed within normal operational behavior. Addressing these threats demands more than static rule-based systems; it calls for adaptive strategies capable of identifying subtle behavioral deviations. This has driven the integration of AI and behavioral analytics to detect anomalies in access logs and system activity.

Recent literature highlights the relevance of such approaches. Aiken et al. [133] introduced Haywire, a robust defense lifecycle for FL that detects poisoned updates by leveraging kernel PCA and clustering strategies. Although the work focuses on maintaining model integrity across siloed, non-i.i.d. environments, the detection of behavioral outliers in decentralized updates has significant implications for identifying insider behavior in collaborative FL-based CTI systems. Ho et al. [134] developed Hopper, a specification-based system that utilizes graph modeling to detect lateral movements in enterprise networks solely based on authentication logs. Hopper identifies adversarial pivoting by tracing causally coherent paths that involve credential switches and access to new hosts—behaviors often associated with insider misuse. Its capability to operate without host-based instrumentation enhances its value for real-time CTI lifecycles.

In a different vein, Pan et al. [135] proposed ASSET, a generalized method for detecting backdoor data across supervised, self-supervised, and transfer learning paradigms. By leveraging nested optimization and adaptive thresholding, ASSET provides robustness against stealthy data poisoning, even in label-sparse environments, making it well-suited for detecting malicious internal data manipulation in CTI systems that rely on unlabeled logs. Zhao et al. [136] conducted a decade-long analysis of malware reports, uncovering structural patterns of malicious infrastructure via co-occurrence modeling and churn metrics. Though not focused on behavioral embeddings, their classification lifecycle and infrastructure-level insights are valuable for identifying persistent insider threat vectors operating across known malicious infrastructures. Fuller et al. [137] presented C3PO, a covert infiltration and measurement framework targeting malware C&C infrastructures. Their observation that over-permissioned protocols can be exploited for access resonates with the insider threat context, wherein privileged users may leverage legitimate but excessive permissions to mask exfiltration. However, their primary contribution lies in extracting malware capabilities, not behavioral modeling.

Zhu et al. [138] proposed ClickScanner, a static detection mechanism for humanoid click fraud based on data dependency graphs and VAE-based anomaly detection. Although the work targets mobile ad fraud, its unsupervised nature

TABLE 4. Key Lessons on NLP for Social Engineering Detection in CTI

Challenge Area	Insight from NLP/LLM Integration	Implication for CTI
Linguistic Manipulation Detection	NLP enables detection of coercive or deceptive patterns (e.g., sentiment, politeness, rhetorical structure)	Enhances phishing and pretexting detection beyond keyword-based filters
Adversarial Language Inputs	LMs are vulnerable to semantic triggers, homographs, and prompt injection	Requires adversarial testing and semantic anomaly detectors in CTI workflows
Latent Model Backdoors	Backdoors can be embedded during pretraining without labeled data	Demands rigorous validation and model provenance tracing
Information Leakage	Embeddings may reveal sensitive entities or threat data	Calls for privacy-preserving embedding strategies in CTI platforms
Synthetic Scenario Generation	NLP can generate realistic adversarial dialogues for training and red-teaming	Supports proactive defense via simulation of social engineering campaigns
Behavioral Intent Modeling	Language analysis allows inference of malicious intent through discourse context	Enables behavior-centric detection models in CTI lifecycles
LLM Exploitability	Generative models may unintentionally expose CTI logic or policies	Highlights the need for output controls and fine-tuned prompt engineering

and bytecode-level feature modeling offer transferable techniques for detecting insider manipulation of telemetry or client-side logic in mobile CTI contexts. Salem et al. [139] introduced dynamic backdoor attacks that circumvent static trigger-based defenses. While their core focus is adversarial ML, the notion of dynamic and context-specific behavior generation aligns with sophisticated insider attacks that evolve over time. Their generative backdoor approach could inform how insiders inject logic into collaborative lifecycles, evading detection by current defenses.

Yu et al. [140] presented FeatureFool, a model extraction attack on commercial MLaaS platforms using adversarial querying and transfer learning. While their research targets black-box model stealing, its principles extend to CTI, where insiders might replicate detection logic to bypass controls. Nevertheless, the study does not directly address behavioral baselining or insider modeling. Thirumuruganathan et al. [141] introduced SIRAJ, a self-supervised framework for aggregating heterogeneous CTI reports into robust threat embeddings. By learning from unlabeled telemetry, SIRAJ enables early detection of threats without dependency on labeled datasets—a characteristic essential for detecting insider activity where ground truth is scarce. However, its core contribution lies in entity classification across CTI sources, not behavioral anomaly detection.

Finally, Deng et al. [142] proposed ARES, a transformer-based website fingerprinting attack model for multi-tab Tor scenarios. While not aimed at insider detection, ARES demonstrates how local pattern extraction and dependency modeling can operate effectively under obfuscation and noise conditions often present in insider contexts. Still, its focus is on external traffic de-anonymization rather than internal misuse classification.

These contributions illustrate the feasibility and relevance of integrating behavioral analysis and AI to detect insider threats. By analyzing usage patterns and anomalies, ML algorithms reveal subtle signs of insider activity. These models facilitate ongoing learning and adaptation to changes in behavior, without requiring strict rules. However, significant obstacles remain, including the lack of high-quality, labeled datasets for training supervised models. Insider threats often generate weak signals amid legitimate activities, requiring models that operate in uncertain environments. Additionally, establishing robust behavioral baselines that account for variations in user roles is crucial. Table 4 presents the key findings from studies on utilizing NLP to detect social engineering attacks in CTI. It presents lessons, effective methods, and practical insights from the research.

H. FL FOR PRIVACY-PRESERVING THREAT INTELLIGENCE SHARING

As cyber threats become more complex and frequent, organizations are urged to share actionable CTI to strengthen their collective defense posture. However, this collaboration is often hindered by concerns over data privacy, operational secrecy, and compliance. FL emerges as a promising paradigm for privacy-preserving CTI sharing by enabling decentralized model training across distributed nodes without the need to share raw telemetry data.

Unlike traditional centralized approaches, FL distributes the training process to the data holders and aggregates only the model updates on a central or decentralized coordinator. This architecture inherently supports privacy, yet it is not without vulnerabilities. Research by Ma et al. [143] reveals that FL is susceptible to Membership Inference Attacks (MIA), where adversaries may infer whether specific data were part of the training set. Their defense framework enables client-side filtering of poisoned updates using semantic distance metrics and gradient confidence monitoring, effectively neutralizing MIAs while preserving model utility. This aligns directly with CTI use cases, where exposure of training samples could equate to leaking sensitive threat indicators or organizational behavior patterns.

Decentralized Learning (DL) may appear to be a logical extension of FL, removing the central server altogether. However, Pasquini et al. [144] demonstrate that DL introduces more severe privacy risks, including gradient inversion and backdoor attacks, due to the increased transparency and spar-

sity of peer-to-peer update exchanges. This indicates that for CTI scenarios—especially those involving highly sensitive or unevenly distributed data—FL remains the superior approach under current threat models.

Incorporating LLMs into FL-based CTI platforms offers powerful capabilities for textual threat analysis, including automated threat report generation and entity extraction. However, LLMs are prone to privacy leakage via unintended memorization, as shown by CodexLeaks [145]. The authors demonstrate how LLMs trained on sensitive codebases can regurgitate confidential data, even under verbatim blocking techniques. This has profound implications for federated CTI lifecycles that fine-tune LLMs on security logs or scripts. Thus, auditing lifecycles and inference control mechanisms must be integrated to prevent leakage via textual outputs.

The importance of anonymizing structural threat data is further underscored by Gao and Liu [146], who propose dual-objective anonymization techniques for graph data such as organizational knowledge graphs or social network-based threat maps. Their approach ensures that node-level privacy (individual entities) is preserved while retaining utility at the subgraph level (group behaviors), a trade-off highly relevant to CTI analytics using FL over network telemetry or attack graphs.

When operationalizing FL in real-world threat detection scenarios, frameworks such as Cerberus [147] illustrate the feasibility of training recurrent models on telemetry from thousands of organizations. While accuracy may decline slightly compared to centralized baselines, the preservation of privacy and the introduction of metrics, such as participant contribution impact, enable a strategic evaluation of data utility and trust. Crucially, this supports fair and scalable collaboration in federated CTI ecosystems. Nevertheless, the assumption that model updates alone are safe for sharing is challenged by Shen et al. [148]. Their framework demonstrates that even node embeddings, when detached, can reveal structural relationships, highlighting the need for stricter embedding sanitization or secure protocols when sharing intermediate representations in FL-driven CTI systems.

Defensive innovations such as DeepSight [149] and FLTrust [150] further demonstrate how FL can be hardened against sophisticated backdoor and poisoning attacks. DeepSight introduces unsupervised clustering techniques to detect anomalous update signatures, while FLTrust uses a curated root dataset to weight and verify client updates. These mechanisms are particularly effective in adversarial CTI environments, where malicious actors may attempt to pollute collective intelligence through compromised endpoints.

Meanwhile, the privacy guarantees of FL must be rigorously accounted for. Yu et al. [151] highlight discrepancies between theoretical differential privacy assumptions and practical SGD batching implementations. Their refined privacy accountant and adaptive noise scheduling strategies ensure that FL models maintain both accuracy and formal privacy guarantees—an essential requirement for CTI platforms handling sensitive cross-border data.

Finally, the issue of data deletion in compliance-driven settings is a non-trivial matter. Chen et al. [152] show that machine unlearning may paradoxically expose training data via differential inference. This presents a risk in CTI deployments where entities may need to revoke previously shared contributions without undermining the platform's overall integrity.

V. ANALYSIS, TRENDS, AND CROSS-CUTTING INSIGHTS ON LLM-DRIVEN CTI

Identifying CTI research opportunities is not just about grouping similar topics or chasing the latest trends in the literature; it also involves understanding the underlying principles and identifying key areas of focus. It is trickier than that. What helps is looking at things from multiple angles—asking not just what is interesting, but also what is useful, how mature the idea is, and whether it stands a chance of fitting into existing systems. With that in mind, we approached each opportunity through three main lenses: where it fits in the CTI lifecycle, how far along it appears in terms of scientific development, and how feasible (or not) it might be to integrate it into a real-world CTI architecture, especially one that relies on LLMs.

What follows is a kind of summary—a consolidated view of what we are calling the ''research diamonds.'' These are opportunities that stood out not just because they were novel or promising, but because they scored reasonably well across those three dimensions. We tried to keep things grounded: for each case, we examined how it aligns with specific stages of the CTI lifecycle, where it stands in terms of maturity (some are still relatively early-stage), and whether it might work as part of a modular LLM-driven CTI system.

Table 5 pulls these findings together. It is not a definitive list, of course. However, it does offer a structured approach to where research might go next—and, perhaps more importantly, where it can go without getting too far ahead of what is technically or operationally realistic.

To understand the trajectory, challenges, and evolution of CTI emerging from the contemporary cyber world, one must not only examine and guide isolated technical innovations but also identify and trace the emerging conceptual and methodological patterns that connect them. The spectral analyses developed throughout this study, particularly through the research diamonds, reveal a phenomenon that transcends any individual contribution: the progressive formation of an integrated paradigm in which LLM and NLP techniques reconfigure the very architecture of cyber defense.

The ubiquity of embedding models and knowledge graphs, for example, indicates a paradigmatic shift towards a CTI that not only consumes indicators but understands and reasons about the semantic relationships between them. Likewise, the rise of FL and model-based security technologies, such as ''watermarking'', directly responds to the need for collaboration in a global threat ecosystem, where trust and privacy are as crucial as detection effectiveness. Accordingly, securing AI infrastructure for collaborative CTI will require sustained investment and deeper study of model watermarking and prove-

TABLE 5. Consolidated Evaluation of Cross-Domain Research Opportunities in LLM-Enhanced CTI

Opportunity (Diamond)	CTI lifecycle phase(s)	Maturity	Impact	Subsystem	Representative Works
Explainability of LLM-based CTI Reasoning	Analysis, Prioritization	Low	High	Reasoning	[80], [153]–[155]
Provenance and Auditability of LLM Outputs	Dissemination, Governance	Low	High	Governance	[64]–[66], [156]–[158]
Federated and Privacy-Preserving Inference for CTI	Collection, Processing	Medium	High	Perception/Governance	[72], [159]–[162]
Multimodal Threat Actor Profiling	Enrichment, Analysis	Medium	High	Reasoning	[69], [163]–[166]
LLM Robustness Against Adversarial CTI Inputs	Collection, Enrichment, Analysis	Low	Medium	Perception/Reasoning	[167]–[182]
Semantic Extraction and Classification from OSINT	Collection, Enrichment	High	High	Perception	[1], [55], [104], [110], [111], [113], [141], [183]–[185]
Secure Neural Architectures for Embedded CTI	Collection, Analysis	Medium	Medium	Reasoning	[186]–[190]
Automated Alignment with CTI Standards (MITRE, NIST)	Dissemination, Governance	Medium	High	Governance	[56], [67], [110], [191]–[193]
LLM-Based Detection of Influence Operations	Enrichment, Analysis	Medium	Medium	Reasoning	[44], [45], [194]–[196]
Anonymization and Redaction of Sensitive CTI	Enrichment, Dissemination	Low	Medium	Governance	[197]–[200]

nance mechanisms to verify model ownership and integrity across federated and shared deployments — e.g., model/IP watermarking (Zhang et al. [42]), LLM/API watermarking (Li et al. [201]), and integrity mechanisms for federated learning (Chowdhury et al. [202]; Lycklama et al. [203]). These technological and methodological pillars collectively form the basis for the development of the quantitative and qualitative validation capabilities discussed below.

LLMs face challenges regarding generalization and operational reliability in SOC environments. A core issue is that, when deployed in detection or triage tasks, LLMs tend to generate a high rate of false positives, often flagging benign activity as suspicious due to their reliance on local, stateless analysis and limited context. To improve precision, it is essential to maintain a persistently updated context that accumulates evidence over time and across multiple sources. However, LLMs remain fundamentally black-box models, susceptible to unpredictable behaviors, adversarial inputs, and failures in rare or critical scenarios. As a result, continuous supervision either by human analysts or by layered, specialized AI agents is indispensable, especially in high-stakes settings.

Despite classic metrics (such as F1 in extraction, classification accuracy, feed coverage/latency), CTI lacks a dedicated benchmark. Therefore, comparisons remain fragmented and sensitive to dataset selection. The path lies in a taxonomy aligned with CTI workflows and metrics that combine effectiveness, calibration/uncertainty, resilience to distribution changes, efficiency, and end-to-end latency, as well as audit trails and compliance (ATT&CK/NIST/ISO). Initial evidence suggests feasibility, but also risks: noisy labels, outdated IOCs, and vendor lock-in distort results; unrealistic protocols tend to inflate performance [64]–[66]. Rather than a single "gold standard," scenario-based suites may offer more honest comparisons.

This transformation is a response to persistent operational challenges in CTI workflows, including the overwhelming volume of alerts, the cognitive fatigue of analysts, the increasing complexity of malware ecosystems, and the growing demand for explainability, compliance, and trust. These pressures have catalyzed a search not only for more powerful tools, as reflected in several initiatives presented in this research, but also for intelligent agents capable of incorporating domain knowledge, learning from dynamic environments, and supporting the human decision-maker with semantic understanding and contextual reasoning.

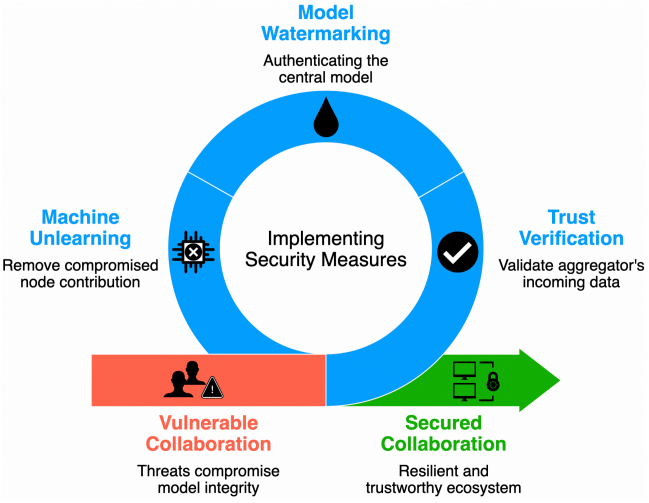

FIGURE 11. CTI of the LLM-assisted lifecycle.

The analytical synthesis of this literature reveals a transition from a reactive and manually intensive paradigm to a proactive, automated, and cognitively augmented ecosystem, orchestrated by intelligent agents. Such agents are structured under three main capabilities: triage, investigation, and mitigation, based and anchored by international bodies such as the Tallinn Manual[3] (Rules 4 and 69), National Institute of Standards and Technology (NIST) SP 800-61r3[4], and ISO/IEC 27035[5]. An interdependent lifecycle of CTI emerges

[3]https://ccdcoe.org/research/tallinn-manual
[4]https://csrc.nist.gov/pubs/sp/800/61/r3/final
[5]https://www.iso.org/standard/78973.html

that is cohesive and expanded by LLM and, above all, more productive. As illustrated by Figure 11. This transformation does not merely represent a gain in efficiency but an essential strategic adaptation to face the speed, scale, and complexity of contemporary threats, articulating a paradigmatic change in the way cyber defense is conceptualized and operationalized.

At the heart of this shift is the role of LLM-driven agents as cognitive and procedural co-pilots in CTI workflows. These agents perform three main functions: triage, investigation, and mitigation. In triage, LLMs act as intermediaries between Security Information and Event Management (SIEM) systems and human analysts, performing semantic correlation, noise suppression, and prioritization of events. In an investigation, they operate as contextual reasoners, mapping IOCs to known Tactics, Techniques, and Procedures (TTPs), synthesizing incident timelines, and querying threat intelligence repositories. Finally, in mitigation, LLMs assist in containment and remediation actions, either by recommending standard operating procedures extracted from security playbooks or by autonomously triggering predefined workflows in SOAR platforms. These capabilities anchor a reconfiguration of the incident response cycle, aligning it more closely with regulatory mandates of explainability, audibility, and proportionality. LLMs may inadvertently leak sensitive training data, raising privacy and compliance concerns. Advances in federated learning and differential privacy [143], [161] offer practical mitigation, but deployment at CTI scale remains challenging. Continued research should prioritize privacy auditing and robust data anonymization throughout the CTI pipeline.

Consider the triage layer first. Security Operations Centers routinely receive millions of low-level alerts, many of which are redundant, irrelevant, or lacking in actionable context. Traditional SIEM systems, while effective at aggregation and correlation, often fall short in prioritization and semantic disambiguation. In this case, LLMs can act as first-line interpreters, filtering and grouping alerts based on threat relevance, narrative coherence, or similarity to known TTPs. Unlike previous rule-based systems, these models adapt to new threat patterns and learn contextual nuances. This capability becomes particularly relevant when dealing with linguistic manipulation and adversarial signal injection, as discussed in [204]. In adversarial settings, LLM-driven CTI pipelines remain vulnerable to crafted inputs and distribution shifts; hardening calls for Unicode-aware canonicalization, adversarial detection, provenance controls, and continuous red-teaming/monitoring across ingestion, triage, and dissemination.

Scaling LLMs for real-time CTI analytics and for (semi)autonomous SOC pipelines requires careful management of computational cost, end-to-end inference latency, and context adaptation under streaming workloads. Recent studies [78], [205] show that model compression, distributed/edge inference, and incremental learning can mitigate these constraints; yet, achieving operations-grade scalability remains challenging given synchronization, drift, and privacy/interpretability requirements. In practice, auto-SOC designs must favor budgeted cascades/routing, streaming retrieval with sliding context, and telemetry-aware batching to sustain triage–investigation–mitigation at scale.

Beyond triage, we enter the realm of investigation. Once alerts are deemed relevant, analysts face the daunting task of reconstructing potential attack paths, linking disparate indicators, and attributing behaviors to actors or campaigns. This phase requires not only data processing but also contextual understanding—a domain in which LLMs excel. Drawing inspiration from work such as [189] and [197], we envision LLMs capable of synthesizing insights into program structures, behavioral graphs, and language-based threat reports, serving as cognitive prosthetics to alleviate human overhead. Fundamentally, these agents do not replace analysts; instead, they extend their reach, enabling deeper forensic analysis and more accurate attribution.

Despite their impressive analytical capacity, Large Language Models are prone to a persistent and often underestimated limitation—hallucinations. These occur when models generate syntactically coherent yet factually inaccurate statements, a behavior that poses critical risks in CTI workflows where accuracy and contextual validity are paramount. Within automated triage or investigation processes, a hallucinated association between threat indicators or actor behaviors could distort incident analysis or propagate misleading intelligence across systems. Recent work by Siino and Tinnirello [206] demonstrates that prompt engineering and few-shot learning can significantly mitigate such occurrences. Their experiments with the Mistral 7B model showed that contextual prompts embedding example-driven reasoning improved hallucination detection, achieving F1-scores of 0.72 and 0.75 on English and Swedish datasets, respectively. These findings suggest that incorporating prompt-based self-verification mechanisms—in which the model is explicitly instructed to reassess or fact-check its own outputs—can substantially enhance the reliability and transparency of LLM-driven CTI systems. In practice, prompt design becomes a defensive layer complementing model security, ensuring that automation augments, rather than undermines, human analytical judgment.

The third layer—mitigation—is perhaps the most delicate. Here, the risk of automating incorrect actions can have real-world consequences. However, with the proper safeguards in place, LLMs can contribute by suggesting countermeasures, generating response scripts, or executing predefined playbooks in collaboration with SOAR systems. The challenge lies in ensuring that such actions remain explainable, auditable, and compliant, especially given the evolving legal landscape and the potential for misattribution or overreaction, as warned in Rules 4 and 69 of the Tallinn Manual. However, what enables this triage-investigation-mitigation continuum is not just the presence of LLMs, but the infrastructure upon which they operate: the security playbook. Traditionally static and textual, the playbook has now evolved into a dynamic, semantically structured knowledge base—a source from which

agents extract rules, validate actions, and infer security policies. Illustrated by Figure 12. Studies such as [207] and [188] provide models for this transformation, demonstrating how NLP can extract actionable constraints from documentation and specifications, converting natural language into machine-verifiable logic. In this sense, the manual ceases to be a reference and becomes an operational substrate, and LLMs become its interpreters.

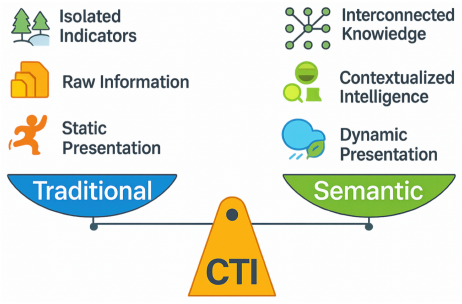

FIGURE 12. **Tradeoff between traditional and semantic approaches to CTI**

It is at this point that the cross-cutting theme becomes apparent. Across domains—from device security [207], firmware compliance [188], code vulnerability detection [189], and even stripped-down binary disassembly [186]—a familiar pattern emerges: the use of language-based models to extract, formalize, and operationalize latent knowledge embedded in text or code. Whether configuration files, hardware manuals, or user reviews [208], they form the linguistic framework from which CTI actors build their understanding of the threat landscape.

This cross-pollination between NLP and CTI creates a methodological bridge: where NLPs not only reason about structured indicators, but also learn from unstructured sources, translate specifications into constraints, and expose vulnerabilities through semantic abstraction, as in GAN-based gadget generation in [177] or cross-architecture code comparison in [187].

What we are witnessing is a shift in epistemology: from reactive detection to proactive reasoning; from signature matching to semantic inference. Moreover, this shift, while technologically enabled, is conceptually anchored in the recognition that cybersecurity is no longer just a technical problem—it is a language problem, a knowledge problem, and ultimately a trust problem.

However, for these systems to mature beyond prototypes, evaluation becomes critical. As argued by [205], many ML-based detection systems lack validation frameworks, especially when they incorporate weak supervision or semi-supervised learning. This insight is even more critical when integrating LLMs, whose probabilistic outcomes and general-purpose nature make them difficult to validate deterministically. A way forward involves not only benchmarking but also the adoption of cost models, transparency protocols, and reproducible methodologies to assess the reliability, utility, and ethical boundaries of LLM-enhanced CTI. Additionally, the effectiveness of this ecosystem of agents depends on its ability to continuously learn from data from multiple sources, often distributed across different organizations. FL has emerged as the leading approach to enabling this collaboration while preserving privacy. However, this introduces a paradox: the collaboration mechanisms themselves become a new and sophisticated attack surface. The security of AI models, and not just data, is emerging as the critical new frontier in cybersecurity. In the literature, we have identified that FL, although promising for its potential to preserve the privacy of sensitive data, has become a fertile ground for sophisticated attacks. The inaugural threat was established by Fang et al. [209], who demonstrated that poisoning attacks, even with normatively limited magnitude, can evade robust aggregators and induce severe degradations in the accuracy of global models. This finding was refined by Shejwalkar et al. [210], who showed that, although production-scale FL environments are more resilient than suggested in academic simulations, they still remain vulnerable to targeted attacks, especially when there is random selection of clients in each round. More worrying is the finding that the offensive not only persists, but evolves: the 3DFed framework [211] introduces an adaptive model that integrates hidden feedback channels (indicators) to evade defenses even in configurations with multiple protection mechanisms.

In parallel, we must recognize that CTI systems do not operate in isolation. They are embedded in organizational routines, inter-institutional exchanges, and legal mandates. Therefore, any proposal for LLM-based agents must include safeguards to ensure that actions align with the NIST incident response lifecycle (preparation, detection, analysis, containment, eradication, recovery, and post-incident treatment) and the ISO/IEC 27035 principles of timeliness, documentation, and legal compliance. Failure to do so could turn technical innovation into strategic liability—especially in international contexts where attribution, escalation, and retaliation are governed by international cyber norms. The inherent uncertainty in real-world data further complicates the task. The work of [205] serves as an important cautionary tale, demonstrating through an evaluation framework that applying semi-supervised learning methods to unlabeled data is not always beneficial and can, in some cases, be detrimental. This highlights the importance of caution and validation before deploying LLM agents that rely on weakly supervised data. However, the attack surface is not purely technical; it is also linguistically oriented. As demonstrated by [204], adversaries can exploit language collisions to poison information ecosystems, such as search engines —a tactic that a CTI agent must also be trained to recognize and mitigate.

This scenario has led to the development of an increasingly refined defense ecosystem, reflecting the severity and complexity of the problem. A key research focus is the analysis of model updates. Mechanisms such as FLARE [212] and FFA [201] shift the analysis from the parameter space to the latent representation space, improving the ability to discriminate malicious updates. Other approaches,

such as BACKDOORMAN [213], use information-theoretic measures (e.g., Kullback-Leibler divergence) to assign reliability scores. Systems like FLAME [214] consolidate defenses in depth, integrating techniques such as update clustering, "adaptive clipping", and differentially private noise to mitigate attacks without compromising accuracy. However, recognizing that statistical defenses are by definition probabilistic and susceptible to evasion, an emerging line of thought seeks formal guarantees of integrity and verifiability. In this sense, solutions such as RoFL [203] and EIFFeL [202] use zero-knowledge proofs and arithmetic circuits to validate that updates satisfy explicit policies, even under encryption, raising the level of trust in systems to an auditable level. In addition, the privacy axis gains density as the risks of leaking sensitive data through unintentional memorization in models become evident, as demonstrated by Béguelin et al. [3]. In this context, initiatives such as GraphEraser [215] emerge, which aim to enable efficient "machine unlearning" while preserving the data structure in graphs.

VI. CONCLUSION

This article presented a survey across the domains of CTI, NLP, and LLMs, aiming to identify strategically relevant research as of six top-tier security conferences over a five-year horizon. Rather than relying on a string of searches or manual screening, we applied LLM-assisted semantic classification to curate a contextually aligned corpus with high thematic fidelity. Enabling the process of 4,651 publications, resulting in the selection and annotation of 287 papers into a systematic review by an LLM-assisted approach.

The review process combines the analytical rigor of traditional systematic surveys with the scalability and contextual awareness provided by language models. In terms of scope and depth, the outcome is comparable to conventional literature reviews, but introduces two novel contributions: (1) a reusable, survey-like dataset that supports rapid information retrieval and can act as a productivity tool for hypothesis generation, literature mapping, and scientific writing; and (2) a semantic synthesis layer that enabled the identification and consolidation of eight strategic and emergent research opportunities, referred to as diamonds, derived from converging themes in the analyzed corpus.

These eight diamonds outline key research paths—from privacy-preserving inference and robust CTI architectures to explainable AI and multimodal integration—poised to advance future cyber threat intelligence systems. Each diamond represents a convergence of emerging technical capabilities and operational CTI needs, informed by recent research across disciplines.

Future work includes operationalizing these findings in real-world CTI environments and releasing our annotated dataset and scripts for reproducibility. The proposed approach is extensible to other security subfields and can help foster strategic foresight in disciplines where technological convergence and knowledge overload pose significant challenges.

ACKNOWLEDGMENT

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001, and by the ITA/PPGAO program. It also received support from the São Paulo Research Foundation (FAPESP), under Grant #2020/09850-0. We thank the C2-DC Laboratory for providing the infrastructure used in the experiments.

REFERENCES

[1] S. Sebastián and J. Caballero, "Towards attribution in mobile markets: Identifying developer account polymorphism," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 771–785. [Online]. Available: https://doi.org/10.1145/3372297.3417281

[2] S. Zhu, Z. Zhang, L. Yang, L. Song, and G. Wang, "Benchmarking label dynamics of virustotal engines," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 2081–2083. [Online]. Available: https://doi.org/10.1145/3372297.3420013

[3] S. Z. Béguelin, L. Wutschitz, S. Tople, V. Rühle, A. Paverd, O. Ohrimenko, B. Köpf, and M. Brockschmidt, "Analyzing information leakage of updates to natural language models," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 363–375. [Online]. Available: https://doi.org/10.1145/3372297.3417880

[4] C. Song and A. Raghunathan, "Information leakage in embedding models," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 377–390. [Online]. Available: https://doi.org/10.1145/3372297.3417270

[5] H. Hu and J. Pang, "Membership inference attacks against gans by leveraging over-representation regions," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2387–2389. [Online]. Available: https://doi.org/10.1145/3460120.3485338

[6] A. Hu, R. Xie, Z. Lu, A. Hu, and M. Xue, "Tablegan-mca: Evaluating membership collisions of gan-synthesized tabular data releasing," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2096–2112. [Online]. Available: https://doi.org/10.1145/3460120.3485251

[7] G. Kaptchuk, T. M. Jois, M. Green, and A. D. Rubin, "Meteor: Cryptographically secure steganography for realistic distributions," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 1529–1548. [Online]. Available: https://doi.org/10.1145/3460120.3484550

[8] A. Alabduljabbar, A. Abusnaina, Ü. Meteriz-Yildiran, and D. Mohaisen, "Automated privacy policy annotation with information highlighting made practical using deep representations," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2378–2380. [Online]. Available: https://doi.org/10.1145/3460120.3485335

[9] D. Bui, Y. Yao, K. G. Shin, J. Choi, and J. Shin, "Consistency analysis of data-usage purposes in mobile apps," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2824–2843. [Online]. Available: https://doi.org/10.1145/3460120.3484536

[10] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the ADAS: securing advanced driver-assistance systems from split-second phantom attacks," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual*

Event, USA, November 9-13, 2020, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 293–308. [Online]. Available: https://doi.org/10.1145/3372297.3423359

[11] H. Kim, J. Bak, K. Cho, and H. Koo, "A transformer-based function symbol name inference model from an assembly language for binary reversing," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, pp. 951–965. [Online]. Available: https://doi.org/10.1145/3579856.3582823

[12] X. Jin, K. Pei, J. Y. Won, and Z. Lin, "Symlm: Predicting function names in stripped binaries via context-sensitive execution-aware code embeddings," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 1631–1645. [Online]. Available: https://doi.org/10.1145/3548606.3560612

[13] Q. Chen, J. Lacomis, E. J. Schwartz, C. L. Goues, G. Neubig, and B. Vasilescu, "Augmenting decompiler output with learned variable names and types," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 4327–4343. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/chen-qibin

[14] X. Li, Y. Qu, and H. Yin, "Palmtree: Learning an assembly language model for instruction embedding," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 3236–3251. [Online]. Available: https://doi.org/10.1145/3460120.3484587

[15] F. Zuo, X. Li, P. Young, L. Luo, Q. Zeng, and Z. Zhang, "Neural machine translation inspired binary code similarity comparison beyond function pairs," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/neural-machine-translation-inspired-binary-code-similarity-comparison-beyond-function-pairs/

[16] Y. Akimoto, K. Fukuchi, Y. Akimoto, and J. Sakuma, "Privformer: Privacy-preserving transformer with MPC," in *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*. IEEE, 2023, pp. 392–410. [Online]. Available: https://doi.org/10.1109/EuroSP57164.2023.00031

[17] M. Du, X. Yue, S. S. M. Chow, T. Wang, C. Huang, and H. Sun, "Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 2665–2679. [Online]. Available: https://doi.org/10.1145/3576915.3616592

[18] Z. Huang, W. Lu, C. Hong, and J. Ding, "Cheetah: Lean and fast secure two-party deep neural network inference," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 809–826. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/huang-zhicong

[19] W. Pei and C. Yue, "Generating content-preserving and semantics-flipping adversarial text," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 975–989. [Online]. Available: https://doi.org/10.1145/3488932.3517397

[20] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible NLP attacks," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 1987–2004. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833641

[21] M. M. Anjum, N. Mohammed, and X. Jiang, "De-identification of unstructured clinical texts from sequence to sequence perspective," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2438–2440. [Online]. Available: https://doi.org/10.1145/3460120.3485354

[22] R. Schuster, T. Schuster, Y. Meri, and V. Shmatikov, "Humpty dumpty: Controlling word meanings via corpus poisoning," in *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 2020, pp. 1295–1313. [Online]. Available: https://doi.org/10.1109/SP40000.2020.00115

[23] J. Li, T. Du, S. Ji, R. Zhang, Q. Lu, M. Yang, and T. Wang, "Textshield: Robust text classification based on multimodal embedding and neural machine translation," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1381–1398. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/li-jinfeng

[24] K. Yuan, H. Lu, X. Liao, and X. Wang, "Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces," in *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, W. Enck and A. P. Felt, Eds. USENIX Association, 2018, pp. 1027–1041. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/yuan-kan

[25] Z. Guo, Z. Lin, P. Li, and K. Chen, "Skillexplorer: Understanding the behavior of skills in large scale," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 2649–2666. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/guo

[26] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-cloak: Intelligibility-, naturalness- & timbre-preserving real-time voice anonymization," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 5181–5198. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/deng-jiangyi-v-cloak

[27] S. Abdelnabi and M. Fritz, "Adversarial watermarking transformer: Towards tracing text provenance with data hiding," in *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*. IEEE, 2021, pp. 121–140. [Online]. Available: https://doi.org/10.1109/SP40001.2021.00083

[28] R. Wu, T. Kim, D. J. Tian, A. Bianchi, and D. Xu, "Dnd: A cross-architecture deep neural network decompiler," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 2135–2152. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/wu-ruoyu

[29] Y. Yuan, S. Wang, and Z. Su, "Precise and generalized robustness certification for neural networks," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 4769–4786. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/yuan-yuanyuan-certification

[30] S. Shen, S. Shinde, S. Ramesh, A. Roychoudhury, and P. Saxena, "Neuro-symbolic execution: Augmenting symbolic execution with neural constraints," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/neuro-symbolic-execution-augmenting-symbolic-execution-with-neural-constraints/

[31] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1613–1630. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179387

[32] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/textbugger-generating-adversarial-text-against-real-world-applications/

[33] L. Pajola and M. Conti, "Fall of giants: How popular text-based mlaas fall against a simple evasion attack," in *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 198–211. [Online]. Available: https://doi.org/10.1109/EuroSP51992.2021.00023

[34] W. M. Si, M. Backes, Y. Zhang, and A. Salem, "Two-in-one: A model hijacking attack against text generation models," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA,*

USA, August 9-11, 2023, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2223–2240. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/si

[35] V. Q. Vo, E. Abbasnejad, and D. C. Ranasinghe, "Ramboattack: A robust and query efficient deep neural network decision exploit," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-239/

[36] R. Zhao, X. Deng, Y. Wang, Z. Yan, Z. Han, L. Chen, Z. Xue, and Y. Wang, "Geesolver: A generic, efficient, and effortless solver with self-supervised learning for breaking text captchas," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1649–1666. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179379

[37] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[38] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 343–362. [Online]. Available: https://doi.org/10.1145/3372297.3417238

[39] J. Chen, J. Wang, T. Peng, Y. Sun, P. Cheng, S. Ji, X. Ma, B. Li, and D. Song, "Copy, right? A testing framework for copyright protection of deep learning models," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 824–841. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833747

[40] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1345–1362. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/jagielski

[41] K. Chen, S. Guo, T. Zhang, X. Xie, and Y. Liu, "Stealing deep reinforcement learning models for fun and profit," in *ASIA CCS '21: ACM Asia Conference on Computer and Communications Security, Virtual Event, Hong Kong, June 7-11, 2021*, J. Cao, M. H. Au, Z. Lin, and M. Yung, Eds. ACM, 2021, pp. 307–319. [Online]. Available: https://doi.org/10.1145/3433210.3453090

[42] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. M. Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018, Incheon, Republic of Korea, June 04-08, 2018*, J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, and T. Kim, Eds. ACM, 2018, pp. 159–172. [Online]. Available: https://doi.org/10.1145/3196494.3196550

[43] H. Harkous, S. T. Peddinti, R. Khandelwal, A. Srivastava, and N. Taft, "Hark: A deep learning system for navigating privacy feedback at scale," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 2469–2486. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833729

[44] S. Prasad, T. Dunlap, A. J. Ross, and B. Reaves, "Diving into robocall content with snorcall," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 427–444. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/prasad

[45] P. Wang, X. Mi, X. Liao, X. Wang, K. Yuan, F. Qian, and R. A. Beyah, "Game of missuggestions: Semantic analysis of search-autocomplete manipulations," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_07A-1_Wang_paper.pdf

[46] I. J. King and H. H. Huang, "Euler: Detecting network lateral movement via scalable temporal graph link prediction," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-227/

[47] Y. Shen, X. He, Y. Han, and Y. Zhang, "Model stealing attacks against inductive graph neural networks," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 1175–1192. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833607

[48] Y. Lyu, Y. Fang, Y. Zhang, Q. Sun, S. Ma, E. Bertino, K. Lu, and J. Li, "Goshawk: Hunting memory corruptions via structure-aware and object-centric memory operation synopsis," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 2096–2113. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833613

[49] D. She, Y. Chen, A. Shah, B. Ray, and S. Jana, "Neutaint: Efficient dynamic taint analysis with neural networks," in *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 2020, pp. 1527–1543. [Online]. Available: https://doi.org/10.1109/SP40000.2020.00022

[50] J. Jang, Y. Lee, A. Kim, B. Na, D. Yhee, B. Lee, J. H. Cheon, and S. Yoon, "Privacy-preserving deep sequential model with matrix homomorphic encryption," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 377–391. [Online]. Available: https://doi.org/10.1145/3488932.3523253

[51] N. Asokan, "Model stealing attacks and defenses: Where are we now?" in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, p. 327. [Online]. Available: https://doi.org/10.1145/3579856.3596441

[52] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Z. Béguelin, "Analyzing leakage of personally identifiable information in language models," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 346–363. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179300

[53] X. Zhang, Z. Zhang, S. Ji, and T. Wang, "Trojaning language models for fun and profit," in *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 179–197. [Online]. Available: https://doi.org/10.1109/EuroSP51992.2021.00022

[54] A. Azizi, I. A. Tahmid, A. Waheed, N. Mangaokar, J. Pu, M. Javed, C. K. Reddy, and B. Viswanath, "T-miner: A generative approach to defend against trojan attacks on dnn-based text classification," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 2255–2272. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/azizi

[55] Y. Qin, Y. Xiao, and X. Liao, "Vulnerability intelligence alignment via masked graph attention networks," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 2202–2216. [Online]. Available: https://doi.org/10.1145/3576915.3616686

[56] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the keyboard? assessing the security of github copilot's code contributions," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 754–768. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833571

[57] Q. Wang, P. Datta, W. Yang, S. Liu, A. Bates, and C. A. Gunter, "Charting the attack surface of trigger-action iot platforms," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, L. Cavallaro, J. Kinder, X. Wang, and J. Katz, Eds. ACM, 2019, pp. 1439–1453. [Online]. Available: https://doi.org/10.1145/3319535.3345662

[58] E. Quiring, D. Arp, and K. Rieck, "Forgotten siblings: Unifying attacks on machine learning and digital watermarking," in *2018 IEEE European Symposium on Security and Privacy, EuroS&P 2018, London, United Kingdom, April 24-26, 2018*. IEEE, 2018, pp. 488–502. [Online]. Available: https://doi.org/10.1109/EuroSP.2018.00041

[59] G. Gómez, P. Moreno-Sanchez, and J. Caballero, "Watch your back: Identifying cybercrime financial relationships in bitcoin through back-and-forth exploration," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou,

C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 1291–1305. [Online]. Available: https://doi.org/10.1145/3548606.3560587

[60] L. Su, X. Shen, X. Du, X. Liao, X. Wang, L. Xing, and B. Liu, "Evil under the sun: Understanding and discovering attacks on ethereum decentralized applications," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 1307–1324. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/su

[61] X. Wang, P. Peng, C. Wang, and G. Wang, "You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018, Incheon, Republic of Korea, June 04-08, 2018*, J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, and T. Kim, Eds. ACM, 2018, pp. 431–442. [Online]. Available: https://doi.org/10.1145/3196494.3196529

[62] S. H. Na, K. Kim, and S. Shin, "Knowledge seeking on the shadow brokers," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, D. Lie, M. Mannan, M. Backes, and X. Wang, Eds. ACM, 2018, pp. 2249–2251. [Online]. Available: https://doi.org/10.1145/3243734.3278512

[63] J. P. Chapman, "SAD THUG: structural anomaly detection for transmissions of high-value information using graphics," in *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, W. Enck and A. P. Felt, Eds. USENIX Association, 2018, pp. 1147–1164. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/chapman

[64] X. Bouwman, V. L. Pochat, P. Foremski, T. van Goethem, C. H. Gañán, G. C. M. Moura, S. Tajalizadehkhoob, W. Joosen, and M. van Eeten, "Helping hands: Measuring the impact of a large threat intelligence sharing community," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 1149–1165. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/bouwman

[65] X. Bouwman, H. Griffioen, J. Egbers, C. Doerr, B. Klievink, and M. van Eeten, "A different cup of ti? the added value of commercial threat intelligence," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 433–450. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/bouwman

[66] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage, "Reading the tea leaves: A comparative analysis of threat intelligence," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 851–867. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/li

[67] H. Y. Chen and S. P. Rao, "On adoptability and use case exploration of threat modeling for mobile communication systems," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2417–2419. [Online]. Available: https://doi.org/10.1145/3460120.3485348

[68] P. H. N. Rajput, P. Rajput, M. Sazos, and M. Maniatakos, "Process-aware cyberattacks for thermal desalination plants," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, AsiaCCS 2019, Auckland, New Zealand, July 09-12, 2019*, S. D. Galbraith, G. Russello, W. Susilo, D. Gollmann, E. Kirda, and Z. Liang, Eds. ACM, 2019, pp. 441–452. [Online]. Available: https://doi.org/10.1145/3321705.3329805

[69] P. Wang, X. Liao, Y. Qin, and X. Wang, "Into the deep web: Understanding e-commerce fraud from autonomous chat with cybercriminals," in *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/into-the-deep-web-understanding-e-commerce-fraud-from-autonomous-chat-with-cybercriminals/

[70] Z. Xi, R. Pang, S. Ji, and T. Wang, "Graph backdoor," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 1523–1540. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/xi

[71] S. Barbieri, "Survey-like dataset of 287 cybersecurity papers selected from top-tier conferences," 2025. [Online]. Available: https://dx.doi.org/10.21227/26gf-8w09

[72] H. Fereidooni, A. Dmitrienko, P. Rieger, M. Miettinen, A. Sadeghi, and F. Madlener, "Fedcri: Federated mobile cyber-risk intelligence," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-229/

[73] G. D. L. T. Parra, L. Selvera, J. Khoury, H. Irizarry, E. Bou-Harb, and P. Rad, "Interpretable federated transformer log learning for cloud threat forensics," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-236/

[74] O. Alrawi, M. Ike, M. Pruett, R. P. Kasturi, S. Barua, T. Hirani, B. Hill, and B. Saltaformaggio, "Forecasting malware capabilities from cyber attack memory images," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 3523–3540. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/alrawi-forecasting

[75] A. Alsaheel, Y. Nan, S. Ma, L. Yu, G. Walkup, Z. B. Celik, X. Zhang, and D. Xu, "ATLAS: A sequence-based learning approach for attack investigation," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 3005–3022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/alsaheel

[76] R. Schuster, C. Song, E. Tromer, and V. Shmatikov, "You autocomplete me: Poisoning vulnerabilities in neural code completion," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 1559–1575. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/schuster

[77] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. N. Venkatakrishnan, "HOLMES: real-time APT detection through correlation of suspicious information flows," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 1137–1152. [Online]. Available: https://doi.org/10.1109/SP.2019.00026

[78] F. Dong, L. Wang, X. Nie, F. Shao, H. Wang, D. Li, X. Luo, and X. Xiao, "DISTDET: A cost-effective distributed cyber threat detection system," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 6575–6592. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/dong-feng

[79] Y. Liu, W. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "ABS: scanning neural networks for back-doors by artificial brain stimulation," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, L. Cavallaro, J. Kinder, X. Wang, and J. Katz, Eds. ACM, 2019, pp. 1265–1282. [Online]. Available: https://doi.org/10.1145/3319535.3363216

[80] R. Pang, Z. Zhang, X. Gao, Z. Xi, S. Ji, P. Cheng, X. Luo, and T. Wang, "Trojanzoo: Towards unified, holistic, and practical evaluation of neural backdoors," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 684–702. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00048

[81] X. Feng, Q. Li, H. Wang, and L. Sun, "Acquisitional rule-based engine for discovering internet-of-thing devices," in *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, W. Enck and A. P. Felt, Eds. USENIX Association, 2018, pp. 327–341. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/feng

[82] T. Cong, X. He, and Y. Zhang, "Sslguard: A watermarking scheme for self-supervised learning pre-trained encoders," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 579–593. [Online]. Available: https://doi.org/10.1145/3548606.3559355

[83] K. Ryan, K. He, G. A. Sullivan, and N. Heninger, "Passive SSH key compromise via lattices," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen,

C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 2886–2900. [Online]. Available: https://doi.org/10.1145/3576915.3616629

[84] S. U. Hussain, M. Javaheripi, M. Samragh, and F. Koushanfar, "COINN: crypto/ml codesign for oblivious inference via neural networks," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 3266–3281. [Online]. Available: https://doi.org/10.1145/3460120.3484797

[85] X. Han, T. F. J. Pasquier, A. Bates, J. Mickens, and M. I. Seltzer, "Unicorn: Runtime provenance-based detector for advanced persistent threats," in *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/unicorn-runtime-provenance-based-detector-for-advanced-persistent-threats/

[86] Y. Chen, Y. Yao, X. Wang, D. Xu, C. Yue, X. Liu, K. Chen, H. Tang, and B. Liu, "Bookworm game: Automatic discovery of LTE vulnerabilities through documentation analysis," in *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*. IEEE, 2021, pp. 1197–1214. [Online]. Available: https://doi.org/10.1109/SP40001.2021.00104

[87] B. Zhao, S. Ji, X. Zhang, Y. Tian, Q. Wang, Y. Pu, C. Lyu, and R. Beyah, "UVSCAN: detecting third-party component usage violations in iot firmware," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 3421–3438. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/zhao-binbin

[88] D. Chang, J. Q. Chen, Z. Li, and X. Li, "Hide and seek: Revisiting dns-based user tracking," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 188–205. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00020

[89] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf

[90] V. Rimmer, D. Preuveneers, M. Juarez, T. van Goethem, and W. Joosen, "Automated website fingerprinting through deep learning," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-1_Rimmer_paper.pdf

[91] M. Siino, F. Giuliano, and I. Tinnirello, "Llm application for knowledge extraction from networking log files," in *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2024, pp. 01–06.

[92] W. U. Hassan, S. Guo, D. Li, Z. Chen, K. Jee, Z. Li, and A. Bates, "Nodoze: Combatting threat alert fatigue with automated provenance triage," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/nodoze-combatting-threat-alert-fatigue-with-automated-provenance-triage/

[93] X. Gong, Y. Chen, W. Yang, Q. Wang, Y. Gu, H. Huang, and C. Shen, "Redeem myself: Purifying backdoors in deep learning models using self attention distillation," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 755–772. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179375

[94] Y. Chen, R. Guan, X. Gong, J. Dong, and M. Xue, "D-DAE: defense-penetrating model extraction attacks," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 382–399. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179406

[95] Z. Zhang, G. Tao, G. Shen, S. An, Q. Xu, Y. Liu, Y. Ye, Y. Wu, and X. Zhang, "PELICAN: exploiting backdoors of naturally trained deep learning models in binary code analysis," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2365–2382. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/zhang-zhuo-pelican

[96] M. Xu, J. Yu, X. Zhang, C. Wang, S. Zhang, H. Wu, and W. Han, "Improving real-world password guessing attacks via bi-directional transformers," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 1001–1018. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/xu-ming

[97] A. Cidon, L. Gavish, I. Bleier, N. Korshun, M. Schweighauser, and A. Tsitkin, "High precision detection of business email compromise," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 1291–1307. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/cidon

[98] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1291–1308. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/salem

[99] V. Battis and A. Penner, "Transformer-based extraction of deep image models," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 320–336. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00028

[100] V. Cochard, D. Pfammatter, C. T. Duong, and M. Humbert, "Investigating graph embedding methods for cross-platform binary code similarity detection," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 60–73. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00012

[101] S. An, Y. Yao, Q. Xu, S. Ma, G. Tao, S. Cheng, K. Zhang, Y. Liu, G. Shen, I. Kelk, and X. Zhang, "Imu: Physical impersonating attack for face recognition system with natural style changes," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 899–916. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179360

[102] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers," *Information Systems*, vol. 121, p. 102342, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437923001783

[103] B. Bhattarai and H. H. Huang, "Steinerlog: Prize collecting the audit logs for threat hunting on enterprise network," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 97–108. [Online]. Available: https://doi.org/10.1145/3488932.3523261

[104] Z. Zhu and T. Dumitras, "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports," in *2018 IEEE European Symposium on Security and Privacy, EuroS&P 2018, London, United Kingdom, April 24-26, 2018*. IEEE, 2018, pp. 458–472. [Online]. Available: https://doi.org/10.1109/EuroSP.2018.00039

[105] H. Shin, W. Shim, J. Moon, J. W. Seo, S. Lee, and Y. H. Hwang, "Cybersecurity event detection with new and re-emerging words," in *ASIA CCS '20: The 15th ACM Asia Conference on Computer and Communications Security, Taipei, Taiwan, October 5-9, 2020*, H. Sun, S. Shieh, G. Gu, and G. Ateniese, Eds. ACM, 2020, pp. 665–678. [Online]. Available: https://doi.org/10.1145/3320269.3384721

[106] A. Kashapov, T. Wu, S. Abuadbba, and C. Rudolph, "Email summarization to assist users in phishing identification," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 1234–1236. [Online]. Available: https://doi.org/10.1145/3488932.3527292

[107] A. Nadeem, S. Verwer, S. Moskal, and S. J. Yang, "Enabling visual analytics via alert-driven attack graphs," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2420–2422. [Online]. Available: https://doi.org/10.1145/3460120.3485361

[108] J. Zeng, X. Wang, J. Liu, Y. Chen, Z. Liang, T. Chua, and Z. L. Chua, "SHADEWATCHER: recommendation-guided cyber threat analysis using system audit records," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE,

2022, pp. 489–506. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833669

[109] D. Hendler, S. Kels, and A. Rubin, "Detecting malicious powershell commands using deep neural networks," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018, Incheon, Republic of Korea, June 04-08, 2018*, J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, and T. Kim, Eds. ACM, 2018, pp. 187–197. [Online]. Available: https://doi.org/10.1145/3196494.3196511

[110] S. M. Milajerdi, B. Eshete, R. Gjomemo, and V. N. Venkatakrishnan, "POIROT: aligning attack behavior with kernel audit records for cyber threat hunting," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, L. Cavallaro, J. Kinder, X. Wang, and J. Katz, Eds. ACM, 2019, pp. 1813–1830. [Online]. Available: https://doi.org/10.1145/3319535.3363217

[111] C. Ryan, S. Dutta, Y. Park, and N. Rastogi, "An ontology-driven knowledge graph for android malware," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2435–2437. [Online]. Available: https://doi.org/10.1145/3460120.3485353

[112] A. van der Heijden and L. Allodi, "Cognitive triaging of phishing attacks," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 1309–1326. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/van-der-heijden

[113] K. Satvat, R. Gjomemo, and V. N. Venkatakrishnan, "Extractor: Extracting attack behavior from threat reports," in *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 598–615. [Online]. Available: https://doi.org/10.1109/EuroSP51992.2021.00046

[114] H. Ding, J. Zhai, Y. Nan, and S. Ma, "AIRTAG: towards automated attack investigation by unsupervised learning with log texts," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 373–390. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/ding-hailun-airtag

[115] E. Altinisik, F. Deniz, and H. T. Sencar, "Provg-searcher: A graph representation learning approach for efficient provenance graph search," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 2247–2261. [Online]. Available: https://doi.org/10.1145/3576915.3623187

[116] F. Charmet, H. C. Tanuwidjaja, T. Morikawa, and T. Takahashi, "Towards polyvalent adversarial attacks on URL classification engines," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 1246–1248. [Online]. Available: https://doi.org/10.1145/3488932.3527282

[117] Z. Lin, K. Xu, C. Fang, H. Zheng, A. A. Jaheezuddin, and J. Shi, "QUDA: query-limited data-free model extraction," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, pp. 913–924. [Online]. Available: https://doi.org/10.1145/3579856.3590336

[118] S. Shan, A. N. Bhagoji, H. Zheng, and B. Y. Zhao, "Poison forensics: Traceback of data poisoning attacks in neural networks," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 3575–3592. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/shan

[119] M. L. Pacheco, M. von Hippel, B. Weintraub, D. Goldwasser, and C. Nita-Rotaru, "Automated attack synthesis by extracting finite state machines from protocol specification documents," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 51–68. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833673

[120] D. Hendler, S. Kels, and A. Rubin, "Amsi-based detection of malicious powershell code using contextual embeddings," in *ASIA CCS '20: The 15th ACM Asia Conference on Computer and Communications Security, Taipei, Taiwan, October 5-9, 2020*, H. Sun, S. Shieh, G. Gu, and G. Ateniese, Eds. ACM, 2020, pp. 679–693. [Online]. Available: https://doi.org/10.1145/3320269.3384742

[121] T. Lv, R. Li, Y. Yang, K. Chen, X. Liao, X. Wang, P. Hu, and L. Xing, "Rtfm! automatic assumption discovery and verification derivation from library document for API misuse detection," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 1837–1852. [Online]. Available: https://doi.org/10.1145/3372297.3423360

[122] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 1138–1156. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833677

[123] X. Zhang, Z. Zhang, Q. Zhong, X. Zheng, Y. Zhang, S. Hu, and L. Y. Zhang, "Masked language model based textual adversarial example detection," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, pp. 925–937. [Online]. Available: https://doi.org/10.1145/3579856.3590339

[124] A. Naseh, K. Krishna, M. Iyyer, and A. Houmansadr, "Stealing the decoding algorithms of language models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 1835–1849. [Online]. Available: https://doi.org/10.1145/3576915.3616652

[125] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on NLP models via linguistic style manipulation," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 3611–3628. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/pan-hidden

[126] E. Bagdasaryan and V. Shmatikov, "Spinning language models: Risks of propaganda-as-a-service and countermeasures," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 769–786. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833572

[127] W. M. Si, M. Backes, J. Blackburn, E. D. Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang, "Why so toxic?: Measuring and triggering toxic behavior in open-domain chatbots," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2659–2673. [Online]. Available: https://doi.org/10.1145/3548606.3560599

[128] K. Moore, C. J. Christopher, D. Liebowitz, S. Nepal, and R. Selvey, "Modelling direct messaging networks with multiple recipients for cyber deception," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 1–19. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00009

[129] S. Li, H. Liu, T. Dong, B. Z. H. Zhao, M. Xue, H. Zhu, and J. Lu, "Hidden backdoors in human-centric language models," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 3123–3140. [Online]. Available: https://doi.org/10.1145/3460120.3484576

[130] L. Shen, S. Ji, X. Zhang, J. Li, J. Chen, J. Shi, C. Fang, J. Yin, and T. Wang, "Backdoor pre-trained models can transfer to all," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 3141–3158. [Online]. Available: https://doi.org/10.1145/3460120.3485370

[131] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 2020, pp. 1314–1331. [Online]. Available: https://doi.org/10.1109/SP40000.2020.00095

[132] Y. Zhang, L. Xu, A. Mendoza, G. Yang, P. Chinprutthiwong, and G. Gu, "Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/life-after-speech-recognition-fuzzing-semantic-misinterpretation-for-voice-assistant-applications/

[133] W. Aiken, P. Branco, and G. Jourdan, "Going haywire: False friends in federated learning and how to find them," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, pp. 593–607. [Online]. Available: https://doi.org/10.1145/3579856.3595790

[134] G. Ho, M. Dhiman, D. Akhawe, V. Paxson, S. Savage, G. M. Voelker, and D. A. Wagner, "Hopper: Modeling and detecting lateral movement," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 3093–3110. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/ho

[135] M. Pan, Y. Zeng, L. Lyu, X. Lin, and R. Jia, "ASSET: robust backdoor data detection across a multiplicity of deep learning paradigms," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2725–2742. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/pan

[136] B. Z. H. Zhao, M. Ikram, H. J. Asghar, M. A. Kâafar, A. Chaabane, and K. Thilakarathna, "A decade of mal-activity reporting: A retrospective analysis of internet malicious activity blacklists," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, AsiaCCS 2019, Auckland, New Zealand, July 09-12, 2019*, S. D. Galbraith, G. Russello, W. Susilo, D. Gollmann, E. Kirda, and Z. Liang, Eds. ACM, 2019, pp. 193–205. [Online]. Available: https://doi.org/10.1145/3321705.3329834

[137] J. Fuller, R. P. Kasturi, A. K. Sikder, H. Xu, B. Arik, V. Verma, E. Asdar, and B. Saltaformaggio, "C3PO: large-scale study of covert monitoring of c&c servers via over-permissioned protocol infiltration," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 3352–3365. [Online]. Available: https://doi.org/10.1145/3460120.3484537

[138] T. Zhu, Y. Meng, H. Hu, X. Zhang, M. Xue, and H. Zhu, "Dissecting click fraud autonomy in the wild," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 271–286. [Online]. Available: https://doi.org/10.1145/3460120.3484546

[139] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 703–718. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00049

[140] H. Yu, K. Yang, T. Zhang, Y. Tsai, T. Ho, and Y. Jin, "Cloudleak: Large-scale deep learning models stealing through adversarial examples," in *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/cloudleak-large-scale-deep-learning-models-stealing-through-adversarial-examples/

[141] S. Thirumuruganathan, M. Nabeel, E. Choo, I. Khalil, and T. Yu, "SIRAJ: A unified framework for aggregation of malicious entity detectors," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 507–521. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833725

[142] X. Deng, Q. Yin, Z. Liu, X. Zhao, Q. Li, M. Xu, K. Xu, and J. Wu, "Robust multi-tab website fingerprinting attacks in the wild," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1005–1022. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179464

[143] M. Ma, Y. Zhang, M. A. P. Chamikara, L. Y. Zhang, M. B. Chhetri, and G. Bai, "Loden: Making every client in federated learning a defender against the poisoning membership inference attacks," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, ASIA CCS 2023, Melbourne, VIC, Australia, July 10-14, 2023*, J. K. Liu, Y. Xiang, S. Nepal, and G. Tsudik, Eds. ACM, 2023, pp. 122–135. [Online]. Available: https://doi.org/10.1145/3579856.3590334

[144] D. Pasquini, M. Raynal, and C. Troncoso, "On the (in)security of peer-to-peer decentralized machine learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 418–436. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179291

[145] L. Niu, M. S. Mirza, Z. Maradni, and C. Pöpper, "Codexleaks: Privacy leaks from code generation language models in github copilot," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2133–2150. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/niu

[146] T. Gao and F. Li, "Machine learning-based online social network privacy preservation," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 467–478. [Online]. Available: https://doi.org/10.1145/3488932.3517405

[147] M. Naseri, Y. Han, E. Mariconti, Y. Shen, G. Stringhini, and E. D. Cristofaro, "CERBERUS: exploring federated prediction of security events," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2337–2351. [Online]. Available: https://doi.org/10.1145/3548606.3560580

[148] Y. Shen, Y. Han, Z. Zhang, M. Chen, T. Yu, M. Backes, Y. Zhang, and G. Stringhini, "Finding MNEMON: reviving memories of node embeddings," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2643–2657. [Online]. Available: https://doi.org/10.1145/3548606.3559358

[149] P. Rieger, T. D. Nguyen, M. Miettinen, and A. Sadeghi, "Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-205/

[150] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," in *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/fltrust-byzantine-robust-federated-learning-via-trust-bootstrapping/

[151] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 332–349. [Online]. Available: https://doi.org/10.1109/SP.2019.00019

[152] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 896–911. [Online]. Available: https://doi.org/10.1145/3460120.3484756

[153] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 2339–2356. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179420

[154] J. Liu, Y. Kang, D. Tang, K. Song, C. Sun, X. Wang, W. Lu, and X. Liu, "Order-disorder: Imitation adversarial attacks for black-box neural ranking models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2025–2039. [Online]. Available: https://doi.org/10.1145/3548606.3560683

[155] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 2339–2356. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179324

[156] Y. Chen, D. Tang, Y. Yao, M. Zha, X. Wang, X. Liu, H. Tang, and B. Liu, "Sherlock on specs: Building LTE conformance tests through automated reasoning," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 3529–3545. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/chen-yi

[157] L. Li, T. Xie, and B. Li, "Sok: Certified robustness for deep neural networks," in *44th IEEE Symposium on Security and Privacy, SP 2023,*

San Francisco, CA, USA, May 21-25, 2023. IEEE, 2023, pp. 1289–1310. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179303

[158] S. Zhang, Y. Cheng, W. Zhu, X. Ji, and W. Xu, "Capatch: Physical adversarial patch against image captioning systems," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 679–696. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/zhang-shibo

[159] Y. Lee, S. Cheon, D. Kim, D. Lee, and H. Kim, "ELASM: error-latency-aware scale management for fully homomorphic encryption," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 4697–4714. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/lee-yongwoo

[160] L. Zhou, Z. Wang, H. Cui, Q. Song, and Y. Yu, "Bicoptor: Two-round secure three-party non-linear computation without preprocessing for privacy-preserving machine learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 534–551. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179449

[161] B. Wang, F. Wu, Y. Long, L. Rimanic, C. Zhang, and B. Li, "Datalens: Scalable privacy preserving training via gradient compression and aggregation," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 2146–2168. [Online]. Available: https://doi.org/10.1145/3460120.3484579

[162] L. K. L. Ng and S. S. M. Chow, "Sok: Cryptographic neural-network computation," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 497–514. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179483

[163] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, and S. Zannettou, "On the evolution of (hateful) memes by means of multimodal contrastive learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 293–310. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179315

[164] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang, "Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 3403–3417. [Online]. Available: https://doi.org/10.1145/3576915.3616679

[165] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 2043–2059. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833644

[166] P. Lv, C. Yue, R. Liang, Y. Yang, S. Zhang, H. Ma, and K. Chen, "A data-free backdoor injection approach in neural networks," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2671–2688. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/lv

[167] Y. Xu and W. Wang, "Demo: Certified robustness on toolformer," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 3673–3675. [Online]. Available: https://doi.org/10.1145/3576915.3624362

[168] J. He and M. T. Vechev, "Large language models for code: Security hardening and adversarial testing," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 1865–1879. [Online]. Available: https://doi.org/10.1145/3576915.3623175

[169] C. Wei, Y. Lee, K. Chen, G. Meng, and P. Lv, "Aliasing backdoor attacks on pre-trained models," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2707–2724. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/wei-chengan

[170] C. Fu, X. Zhang, S. Ji, T. Wang, P. Lin, Y. Feng, and J. Yin, "Freeeagle: Detecting complex neural trojans in data-free cases," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 6399–6416. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/fu-chong

[171] J. Gong, W. Zhang, C. Zhang, and T. Wang, "Surakav: Generating realistic traces for a strong website fingerprinting defense," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 1558–1573. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833722

[172] S. Shan, W. Ding, E. Wenger, H. Zheng, and B. Y. Zhao, "Post-breach recovery: Protection against white-box adversarial examples for leaked DNN models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2611–2625. [Online]. Available: https://doi.org/10.1145/3548606.3560561

[173] Y. Liu, G. Shen, G. Tao, S. An, S. Ma, and X. Zhang, "Piccolo: Exposing complex backdoors in NLP transformer models," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 2025–2042. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833579

[174] N. Mangaokar, J. Pu, P. Bhattacharya, C. K. Reddy, and B. Viswanath, "Jekyll: Attacking medical image diagnostics using deep generative models," in *IEEE European Symposium on Security and Privacy, EuroS&P 2020, Genoa, Italy, September 7-11, 2020*. IEEE, 2020, pp. 139–157. [Online]. Available: https://doi.org/10.1109/EuroSP48549.2020.00017

[175] H. Griffioen, K. Oosthoek, P. van der Knaap, and C. Doerr, "Scan, test, execute: Adversarial tactics in amplification ddos attacks," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds. ACM, 2021, pp. 940–954. [Online]. Available: https://doi.org/10.1145/3460120.3484747

[176] H. Liu, J. Jia, and N. Z. Gong, "Poisonedencoder: Poisoning the unlabeled pre-training data in contrastive learning," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 3629–3645. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/liu-hongbin

[177] M. C. Tol, B. Gülmezoglu, K. Yurtseven, and B. Sunar, "Fastspec: Scalable generation and detection of spectre gadgets using neural embeddings," in *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 616–632. [Online]. Available: https://doi.org/10.1109/EuroSP51992.2021.00047

[178] A. Salem, M. Backes, and Y. Zhang, "Get a model! model hijacking attack against machine learning models," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-241/

[179] J. Zhou, Y. Chen, C. Shen, and Y. Zhang, "Property inference attacks against gans," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-240/

[180] R. Shetty, B. Schiele, and M. Fritz, "A4NT: author attribute anonymity by adversarial training of neural machine translation," in *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, W. Enck and A. P. Felt, Eds. USENIX Association, 2018, pp. 1633–1650. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/shetty

[181] Y. Li, M. Li, B. Luo, Y. Tian, and Q. Xu, "Deepdyve: Dynamic verification for deep neural networks," in *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds. ACM, 2020, pp. 101–112. [Online]. Available: https://doi.org/10.1145/3372297.3423338

[182] S. Li, A. Neupane, S. Paul, C. Song, S. V. Krishnamurthy, A. K. Roy-Chowdhury, and A. Swami, "Stealthy adversarial perturbations against real-time video classification systems," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[Online]. Available: https://www.ndss-symposium.org/ndss-paper/stealthy-adversarial-perturbations-against-real-time-video-classification-systems/

[183] Y. Dong, W. Guo, Y. Chen, X. Xing, Y. Zhang, and G. Wang, "Towards the detection of inconsistencies in public security vulnerability reports," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 869–885. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/dong

[184] X. Shu, F. Araujo, D. L. Schales, M. P. Stoecklin, J. Jang, H. Huang, and J. R. Rao, "Threat intelligence computing," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, D. Lie, M. Mannan, M. Backes, and X. Wang, Eds. ACM, 2018, pp. 1883–1898. [Online]. Available: https://doi.org/10.1145/3243734.3243829

[185] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 2633–2650. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[186] K. Pei, J. Guan, D. Williams-King, J. Yang, and S. Jana, "XDA: accurate, robust disassembly with transfer learning," in *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/xda-accurate-robust-disassembly-with-transfer-learning/

[187] Z. Luo, P. Wang, B. Wang, Y. Tang, W. Xie, X. Zhou, D. Liu, and K. Lu, "Vulhawk: Cross-architecture vulnerability detection with entropy-based binary code search," in *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society, 2023. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/vulhawk-cross-architecture-vulnerability-detection-with-entropy-based-binary-code-search/

[188] W. Zhou, L. Zhang, L. Guan, P. Liu, and Y. Zhang, "What your firmware tells you is not how you should emulate it: A specification-guided approach for firmware emulation," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 3269–3283. [Online]. Available: https://doi.org/10.1145/3548606.3559386

[189] N. T. Islam, G. D. L. T. Parra, D. Manuel, E. Bou-Harb, and P. Najafirad, "An unbiased transformer source code learning with semantic vulnerability graph," in *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*. IEEE, 2023, pp. 144–159. [Online]. Available: https://doi.org/10.1109/EuroSP57164.2023.00018

[190] S. Sajadmanesh, A. S. Shamsabadi, A. Bellet, and D. Gatica-Perez, "GAP: differentially private graph neural networks with aggregation perturbation," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 3223–3240. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/sajadmanesh

[191] R. del Pino, T. Prest, M. Rossi, and M. O. Saarinen, "High-order masking of lattice signatures in quasilinear time," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1168–1185. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179342

[192] Y. Chen, D. Tang, Y. Yao, M. Zha, X. Wang, X. Liu, H. Tang, and D. Zhao, "Seeing the forest for the trees: Understanding security hazards in the 3gpp ecosystem through intelligent analysis on change requests," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 17–34. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/chen-yi

[193] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "Policylint: Investigating internal privacy policy contradictions on google play," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 585–602. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/andow

[194] F. Sharevski, A. Devine, E. Pieroni, and P. Jachim, "Folk models of misinformation on social media," in *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society, 2023. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/folk-models-of-misinformation-on-social-media/

[195] S. Abdelnabi and M. Fritz, "Fact-saboteurs: A taxonomy of evidence manipulation attacks against fact-verification systems," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 6719–6736. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/abdelnabi

[196] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, W. Enck and A. P. Felt, Eds. USENIX Association, 2018, pp. 531–548. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[197] Y. Nan, Z. Yang, X. Wang, Y. Zhang, D. Zhu, and M. Yang, "Finding clues for your secrets: Semantics-driven, learning-based privacy discovery in mobile apps," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-1_Nan_paper.pdf

[198] G. Tao, Q. Xu, Y. Liu, G. Shen, S. An, J. Xu, X. Zhang, and Y. Yao, "MIRROR: model inversion for deep learningnetwork with high fidelity," in *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/auto-draft-203/

[199] P. Santhanam, H. Dang, Z. Shan, and I. Neamtiu, "Scraping sticky leftovers: App user information left on servers after account deletion," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 2145–2160. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833720

[200] D. Bui, B. Tang, and K. G. Shin, "Do opt-outs really opt me out?" in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 425–439. [Online]. Available: https://doi.org/10.1145/3548606.3560574

[201] Z. Li, C. Wang, S. Wang, and C. Gao, "Protecting intellectual property of large language model-based code generation apis via watermarks," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds. ACM, 2023, pp. 2336–2350. [Online]. Available: https://doi.org/10.1145/3576915.3623120

[202] A. R. Chowdhury, C. Guo, S. Jha, and L. van der Maaten, "Eiffel: Ensuring integrity for federated learning," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 2535–2549. [Online]. Available: https://doi.org/10.1145/3548606.3560611

[203] H. Lycklama, L. Burkhalter, A. Viand, N. Küchler, and A. Hithnawi, "Rofl: Robustness of secure federated learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 453–476. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179400

[204] M. Joslin, N. Li, S. Hao, M. Xue, and H. Zhu, "Measuring and analyzing search engine poisoning of linguistic collisions," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 1311–1325. [Online]. Available: https://doi.org/10.1109/SP.2019.00025

[205] G. Apruzzese, P. Laskov, and A. Tastemirova, "Sok: The impact of unlabelled data in cyberthreat detection," in *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*. IEEE, 2022, pp. 20–42. [Online]. Available: https://doi.org/10.1109/EuroSP53844.2022.00010

[206] M. Siino and I. Tinnirello, "Gpt hallucination detection through prompt engineering," in *Proc. of the 25th Working Notes of the Conference and Labs of the Evaluation Forum*, vol. 3740, 2024, pp. 712–721.

[207] Y. Yang, M. Elsabagh, C. Zuo, R. Johnson, A. Stavrou, and Z. Lin, "Detecting and measuring misconfigured manifests in android apps," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 3063–3077. [Online]. Available: https://doi.org/10.1145/3548606.3560607

[208] D. C. Nguyen, E. Derr, M. Backes, and S. Bugiel, "Short text, large effect: Measuring the impact of user reviews on android app security & privacy," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 555–569. [Online]. Available: https://doi.org/10.1109/SP.2019.00012

[209] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1605–1622. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[210] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*. IEEE, 2022, pp. 1354–1371. [Online]. Available: https://doi.org/10.1109/SP46214.2022.9833647

[211] H. Li, Q. Ye, H. Hu, J. Li, L. Wang, C. Fang, and J. Shi, "3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning," in *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 2023, pp. 1893–1907. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179401

[212] N. Wang, Y. Xiao, Y. Chen, Y. Hu, W. Lou, and Y. T. Hou, "FLARE: defending federated learning against model poisoning attacks via latent space representations," in *ASIA CCS '22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, Y. Suga, K. Sakurai, X. Ding, and K. Sako, Eds. ACM, 2022, pp. 946–958. [Online]. Available: https://doi.org/10.1145/3488932.3517395

[213] N. Lukas and F. Kerschbaum, "PTW: pivotal tuning watermarking for pre-trained image generators," in *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, J. A. Calandrino and C. Troncoso, Eds. USENIX Association, 2023, pp. 2241–2258. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/lukas

[214] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A. Sadeghi, and T. Schneider, "FLAME: taming backdoors in federated learning," in *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, K. R. B. Butler and K. Thomas, Eds. USENIX Association, 2022, pp. 1415–1432. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen

[215] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "Graph unlearning," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds. ACM, 2022, pp. 499–513. [Online]. Available: https://doi.org/10.1145/3548606.3559352

SIDNEI BARBIERI received the B.Sc. degree in computer science from the Regional Integrated University of Alto Uruguai and Missões (URI), Santo Ângelo, RS, Brazil, in 2009, and the M.Sc. degree in computer science from the Aeronautics Institute of Technology (ITA), São José dos Campos, SP, Brazil, in 2018.

He is currently pursuing a Ph.D. degree in computer science at ITA and is a visiting researcher at Carnegie Mellon University (CMU), Pittsburgh, PA, USA. Previously, he served as an Officer at the Cyber Defense Center of the Brazilian Army, focusing on CTI, data science, and network security. His research interests include CTI and the application of LLMs to cybersecurity domains.

FLAVIO LUIZ DOS SANTOS DE SOUZA is a Ph.D. candidate in Electrical and Computer Engineering at the Aeronautics Institute of Technology (ITA), Brazil. He is currently a Visiting Researcher at Linköping University (LiU), Sweden, where he develops research on mission-driven communication systems for unmanned aerial vehicles (UAVs) operating in complex and contested environments. His research focuses on integrating satellite links, adaptive communication protocols, and AI-based resource optimization to enhance UAV autonomy, reliability, and mission coordination in both civilian and defense applications.

He holds a Master's degree (M.Sc.) in Computer Science and Computational Mathematics from the University of São Paulo (USP) and a Bachelor's degree (B.Sc.) in Computer Engineering with honors from the University of Araraquara (UNIARA). His academic work includes publications on UAV swarm intelligence, simulation-based mission planning, and capability-based communication modeling. all of which are aligned with military-oriented frameworks such as Capability-Based Planning (CBP).

MARCIO ANDREY TEIXEIRA is a Senior Member of IEEE, and a full professor at Federal Institute of Education, Science and Technology of São Paulo, Campus Catanduva, Brazil. Dr. Marcio received his Ph.D. in Electrical Engineering from the Federal University of Uberlândia, Brazil, in 2012, and his MSc degree in Computer Science from the same university in 2004. He was a postdoctoral researcher in Computer Science and Engineering at Washington University in St. Louis, MO, USA (2017-2018), under the supervision of Professor Dr. Raj Jain, a Life Fellow of the IEEE. His current research includes cybersecurity, network security, machine learning, deep learning, IA for cybersecurity and networking, performance analysis of computer networks, quality of service (QoS) in wireless networks, wireless network protocols, and wireless next generation.

CESAR AUGUSTO CAVALHEIRO MARCONDES received the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA) in 2008.

From 2008 to 2018, he served as an Assistant Professor at the Federal University of São Carlos (UFSCar), where he worked on future Internet architectures and congestion control. He is currently an Assistant Professor in the Department of Computer Systems at the Aeronautics Institute of Technology (ITA), São José dos Campos, SP, Brazil. He also holds patents from the USPTO related to congestion control, obtained during his work with Sun Microsystems. His research interests include computer networks and security.

LOURENÇO ALVES PEREIRA JÚNIOR received the B.Sc. degree (Hons.) in computer science from the University of Alfenas (UNIFENAS) in 2006, and the M.Sc. and Ph.D. degrees in computer science and computational mathematics from the University of São Paulo (ICMC/USP) in 2010 and 2016, respectively.

He is currently an Assistant Professor in the Department of Computer Systems at the Aeronautics Institute of Technology (ITA), São José dos Campos, SP, Brazil. He is a member of the Laboratory of C&C and Cyber Security. His research interests include computer networks and cybersecurity. He also serves as a reviewer for journals such as IEEE Communications Magazine, IEEE Access, IEEE Latin America Transactions, IEEE Transactions on Parallel and Distributed Systems, Journal of Network and Systems Management (JNSM), and Vehicular Communications (Elsevier).