Journal of Cybersecurity and Information Management

♦ ISSN: 2692-4048 Vol: 18, No. 02, 2026 ♦ PP. 01–10

♦ Article DOI: https://doi.org/10.54216/JCIM.180201

FULL LENGTH ARTICLE

An Integrated Framework for Automated Cyber Threat Actor Attribution: LLM-Based Entity Extraction over STIX 2.1 Knowledge Representations with Bayesian Scoring

Aykhan Huseynli[1,*]

[1] Department of Cybersecurity, Azerbaijan Technical University, Azerbaijan

Email: aykhan.huseynli@cert.az

Received: May 04, 2026 Revised: August 07, 2026 Accepted: November 01, 2026 ⋆ Corresponding author

ABSTRACT

Cyber threat attribution has become an essential component of national and organizational cyber defense, enabling analysts to link attacks to known threat actors and inform proportionate responses. However, existing attribution practices remain largely non-reproducible, dependent on implicit analyst judgment, and unable to quantify uncertainty in a principled way. The absence of a common intelligence schema further limits cross-organizational comparison and automation. Through a systematic review of attribution methodologies, this study identifies reproducibility, structured representation, and probabilistic calibration as the central unresolved challenges. To address these gaps, this paper proposes an integrated end-to-end framework combining large language model (LLM)-based named entity recognition, STIX 2.1 knowledge representation, and Bayesian probabilistic scoring. The framework implements a unique combination of Campaign Frequency scoring, Inverse Document Frequency-weighted Jaccard similarity, and feature-class-specific Laplace smoothing that correctly distinguishes durable behavioral evidence from perishable technical indicators. The effectiveness of the proposed pipeline is demonstrated using a manually annotated corpus of 25 real-world CTI reports across 15 entity types. The Bayesian attribution engine, evaluated against known threat actors, produces realistic, analyst-interpretable probability distributions.

Keywords: Cyber Threat Intelligence ▪ Named Entity Recognition ▪ Large Language Models ▪ STIX 2.1 ▪ Threat Actor Attribution ▪ Bayesian Inference

1. INTRODUCTION

The volume of publicly available Cyber Threat Intelligence (CTI) publications has grown substantially over the past decade. Organizations such as Mandiant, IBM X-Force, ESET, Kaspersky Securelist, and Check Point Research routinely publish detailed campaign analyses covering advanced persistent threat (APT) actors, malware families, and attack infrastructure. Attributing a specific cyberattack to a known threat actor is foundational to strategic cyber defense since accurate attribution enables proportionate countermeasures, informs diplomatic and legal responses, and supports long-term defensive posture [1].

Yet the technical attribution process remains largely artisanal. Analysts correlate indicators of compromise (IoCs), behavioral patterns, and contextual clues across heterogeneous sources without a reproducible methodological framework. Three principal deficiencies characterize existing attribution practice. First, results are non-reproducible as they depend on implicit analyst judgment that cannot be audited, challenged, or transferred across organizations. Second, intelligence is rarely represented in a common schema, making

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

cross-organizational comparison prohibitively difficult. Third, probabilistic uncertainty is seldom quantified as outputs are typically binary (attributed or not attributed), which misrepresents the epistemic state of the evidence and can lead to misattribution with serious geopolitical consequences [2]. Additionally, a comprehensive survey of automated APT attribution identifies reproducibility and uncertainty quantification as persistent open problems across the field [3].

Addressing these deficiencies requires three complementary components. The first prerequisite is structured extraction. Before any attribution pipeline can operate, entities must be reliably pulled from unstructured CTI reports. Large language models offer a compelling solution here since their in-context learning capabilities allow entity extraction without task-specific fine-tuning [4, 5]. The second prerequisite is a common representational schema: STIX 2.1 provides a rigorously defined ontology for representing threat actors, malware, tools, infrastructure, vulnerabilities, and their semantic relationships [6], enabling consistent machine-readable feature extraction that is portable across organizations. The third prerequisite is principled uncertainty quantification. Bayesian inference provides a framework for updating belief about actor identity as evidence accumulates, producing communicable probability distributions rather than binary verdicts.

Considering the aforementioned points, this paper makes the following contributions:

(i) An end-to-end attribution pipeline integrating LLM-based NER, STIX 2.1 knowledge representation, and Bayesian probabilistic scoring;

(ii) An LLM-based structured CTI extraction method tested on Claude Sonnet 4.6, GPT-5.4, and LLaMA 4 Scout using annotated CTI reports across 15 entity types and 1,284 ground-truth instances;

(iii) A STIX 2.1 transformation layer with union-based multi-report aggregation and source-confidence filtering, enabling reproducible cross-organizational attribution;

(iv) CF-IDF-weighted Bayesian scoring with feature-class-specific Laplace smoothing that distinguishes artifact evidential value by class, min-max normalization, and a dark figure advisory mechanism evaluated against 17 actors and 57 reports.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 describes the dataset and proposed framework. Section 4 presents experimental results. Section 5 discusses the findings, and Section 6 concludes.

2. RELATED WORK

2.1 The Attribution Problem: Reproducibility, Schema, and Uncertainty

The foundational challenge of cyber threat attribution was framed by Rid and Buchanan [1], who argued that attribution is an art as well as a science. They describe attribution as a nuanced process operating simultaneously at tactical, operational, and strategic levels, dependent on political context and analyst judgment rather than a reproducible technical procedure. This formulation codified a problem rather than solving it, leaving attribution outputs tied to the producing analyst where they cannot be audited, transferred, or challenged. Pahi and Skopik proposed a Cyber Attribution Model (CAM) as a conceptual step toward systematic practice, explicitly noting a lack of a reliable approach that deals with the interdisciplinary challenges of cyber attribution, and highlighting the compounding risk of false flag operations as a fundamental limit on technically derived conclusions [2]. Their companion work catalogued the spoofability of technical artifacts along the kill chain, demonstrating that network-level indicators are easily fabricated while behavioral evidence such as malware families is considerably more durable [7].

Operational evidence confirms these structural weaknesses. Saha et al.'s interviews with 15 security practitioners found that inconsistent naming conventions hinder the merging and correlation of threat information from disparate sources, while tooling and collaboration gaps further complicate attribution [8]. A 2024 master's thesis evaluating a generative-AI tool for threat attribution concluded that automation remains limited by the quality and standardization of available CTI [9]. Two recent comprehensive surveys confirm that these deficiencies remain unresolved at the field level. Rani et al. [3] identify reproducibility and uncertainty quantification as the most persistent open problems across automated APT attribution approaches, noting that prior literature lacks a systematic review of artifacts and methods. Prasad et al. [10] characterize the broader attribution literature as providing fragmented analyses that predate advances in AI/ML and lack a unifying framework. Deka et al. [11] independently confirm that attribution in practice remains a manual process given the complex nature of the task.

2.2 Conceptual Frameworks for Attribution

Several foundational models provide structured vocabularies for attribution reasoning. The Diamond Model decomposed intrusions into adversary, capability, infrastructure, and victim, providing a taxonomy for analyst reasoning [12]. Bahrami et al. [13] extended this into a kill-chain-based taxonomy mapping TTPs to over 40 APT campaigns, useful for understanding adversary behavior but not operationalized as a reproducible scoring pipeline. Li et al. [14] proposed association analysis combining statistical, behavioral, and probability-similarity association over threat intelligence data, while a companion work used a Bayesian network over kill-chain evidence chains to produce probabilistic reasoning about attacker identity [15]. Edie et al. attributed attacks using weighted Jaccard similarity over TTP playbooks extended by association rule mining, achieving 96% accuracy on MITRE ATT&CK data [16]. These frameworks produce useful insights but share three limitations. First, none uses a standardized knowledge representation enabling cross-organizational reuse. Second, none quantifies the differential evidential value of different artifact classes. Third, the probabilistic approaches lack calibration against a diverse open-source actor corpus.

2.3 Machine Learning and Deep Learning Attribution

The dominant recent approach applies supervised ML or deep learning classifiers directly to narrow feature sets. Rosenberg et al. [17] trained end-to-end deep neural networks on malware binary features for nation-state APT attribution, produc-

ing high closed-world accuracy but no uncertainty estimate and requiring proprietary malware samples. Alkaabi and Olatunji applied five classifiers to Amazon Web Services (AWS) honeypot network data, with SVM achieving 94.79% accuracy—a deterministic binary output with no probabilistic calibration [18]. Noor et al. [19] profiled 327 CTI reports using distributional semantics, explicitly noting that CTI is textual and unstructured and cannot be directly interpreted by machines. Their artifact scope is restricted to TTPs, tools, and malware, leaving network-level indicators unaddressed. Böge et al. [20] applied a hybrid transformer-CNN architecture to command-sequence behavioral data, achieving F1 above 95% on high-count datasets but confining attribution evidence to a single behavioral modality. Xiao et al. [21] proposed APT-MMF, fusing attribute, NLP text, and topological graph features through heterogeneous graph attention networks. This is one of the broadest artifact scopes in supervised learning, yet the method produces deterministic class labels without calibrated posterior probabilities. Xiang et al. [22] proposed IPAttributor, restricting evidence exclusively to IPv4 addresses enriched with commercial threat intelligence. All these approaches produce deterministic classification outputs while none quantifies residual uncertainty in the attribution decision.

2.4 Behavioral and Hybrid Feature Attribution

A smaller but important body of work moves beyond purely technical indicators. Irshad and Siddiqui demonstrated that combining technical features (TTPs, tools, target country) with behavioral features (attacker context and motives from Thai CERT's Threat Actor Encyclopedia) improves attribution F1-measure to 95.75% [23], with a subsequent context-aware extension reaching 97% [24]. These results establish the value of expanding the artifact scope to include behavioral evidence. Ali Khan and Almulhem further argued for integrating TTP-based analysis with malware sample analysis to capture the full attack vector, demonstrating the approach on a real Black Basta campaign [25]. However, these approaches extract features from closed, proprietary datasets and apply uniform ML feature weighting that treats all artifact types as equally informative. Neither do they ground their feature space in a standardized interoperability schema, nor do they produce probabilistic outputs. Classification decisions are consequently binary, with no uncertainty communicated to the analyst.

2.5 Probabilistic and Modular Attribution

Two works approach attribution probabilistically and merit specific attention. Li et al. [15] construct a Bayesian network over intrusion kill-chain evidence, producing probabilistic confidence in actor hypotheses. Teuwen proposed a modular architecture using opinion pools that combines outputs from multiple specialized attributors and produces a Probability Mass Function (PMF) over candidate actors, demonstrating that modular probabilistic combination can improve both precision and recall [26]. Both represent meaningful steps toward principled uncertainty quantification. However, neither grounds its knowledge representation in a standardized, interoperable schema, operationalizes differential weighting of artifact classes by their evidential value or perishability, nor provides a mechanism for communicating the structural incompleteness of open-source intelligence, which this work terms the dark figure problem. The analogy is drawn from criminology, where the dark figure of crime refers to the gap between events that occur and those that are observed and recorded. In the attribution context, threat actor profiles constructed from public CTI are necessarily incomplete because campaigns go unreported, vendors selectively disclose findings, and actors deliberately conceal their traces.

2.6 Artifact Taxonomy and Evidential Weighting

The question of which artifacts carry more attribution signal has received limited but important formal attention. Skopik and Pahi qualitatively rated artifact spoofability and relevance along the kill chain, demonstrating that network-level indicators are highly perishable while malware families and TTPs are more actor-specific [7]. Hwang and Kim advanced this through a formal expert elicitation study using Delphi and AHP methods to rank 25 artifact types by evidential value in the presence of false flag operations, providing empirical validation that different artifact classes carry fundamentally different attribution weight [27]. Critically, both studies stop at the analytical level, as neither translates the artifact ranking into an operational scoring pipeline. To our knowledge, no prior automated attribution system operationalizes differential evidential weighting across artifact classes by assigning distinct base weights, IDF-adjusted rarity scores, campaign frequency consistency measures, and class-specific Laplace smoothing floors within a unified pipeline grounded in a standardized knowledge schema.

2.7 Knowledge Graph and STIX-Based Attribution

Mavroeidis and Bromander established STIX 2.1 as the dominant CTI interoperability standard, demonstrating its advantages over alternative taxonomies for enabling automation [6]. Bromander et al. [28] proposed a strict STIX-based data model that enables deduction of new knowledge from shared intelligence, arguing that standardization is a prerequisite for scalable attribution. Marchiori et al. [29] demonstrated automated extraction of STIX objects from CTI narrative reports, enabling scalable pipeline construction. Ren et al. [30] constructed CSKG4APT, a cybersecurity knowledge graph from open-source APT reports combining TTPs, malware, and infrastructure, though attribution is performed through graph matching without probabilistic scoring or uncertainty communication. King et al. [31] proposed TRAIL, a knowledge graph of IoC co-occurrences that achieves 84% attribution accuracy using graph neural networks, representing a strong graph-based result in the literature, but the approach is restricted to network-level IoCs and produces deterministic outputs.

2.8 CTI Named Entity Recognition

Automated extraction of structured entities from CTI text is the prerequisite for any pipeline-based attribution approach. Evangelatos et al. [32] established a transformer-based NER baseline on the Dataset for NER in Threat Intelligence (DNRTI), demonstrating that fine-tuned models substantially outperform rule-based predecessors. Zhou et al. [33] paired BERT with whole-word masking and a BiLSTM-CRF head, improving precision on cybersecurity-specific entities, while Chen et al. [34] showed that a simplified BERT-CRF (omit-

ting the BiLSTM) is competitive and directly comparable to GPT-3.5 in prompt-based inference. Domain-specific pretraining has become a prerequisite for competitive performance: CyBERT [35] and SecureBERT 2.0 [36] demonstrate consistent gains over general-purpose models on downstream CTI tasks. On the benchmark side, Ech-Chammakhy et al. [37] harmonized four CTI NER datasets onto the STIX 2.1 standard, achieving approximately 30% relative F1 improvement over naive concatenation baselines through schema unification. Chan et al. [38] introduced FeedRef2022 derived from 1,854 threat intelligence reports for IoC extraction evaluation, and Della Penna et al. [39] introduced CTI-HAL, a manually annotated dataset structured according to MITRE ATT&CK with inter-annotator agreement validated using Krippendorff's alpha.

2.9 Summary of Key Limitations

Table 1 synthesizes the principal limitations of representative approaches across the five dimensions most relevant to the deficiencies identified above. Three observations emerge from the synthesis. First, the overwhelming majority of approaches restrict their artifact scope to a single modality and treat all features within that scope as equally weighted; no prior automated system operationalizes the artifact ranking established by Hwang and Kim [27] within a unified scoring pipeline. Second, most methods produce deterministic classification outputs; the few probabilistic exceptions (Li et al. [15], Teuwen [26]) lack a standardized knowledge schema that would enable cross-organizational comparison. Third, no existing system combines the four properties simultaneously, namely broad multi-class artifact coverage, principled artifact weighting, a STIX 2.1 standardized representation, and calibrated probabilistic output with explicit uncertainty communication.

3. MATERIAL AND METHOD

3.1 Dataset

The NER evaluation corpus comprises 25 publicly available CTI reports drawn from Mandiant, IBM X-Force, ESET, Kaspersky Securelist, Check Point Research, Trend Micro, Deep Instinct, and Sekoia. Reports were collected from the APTnotes dataset, which indexes threat intelligence publications up to 2024, and Malpedia [40], from which additional reports cover the 2025–2026 period. Reports range from 4,466 to 24,348 tokens, with a median of approximately 8,532 tokens.

The corpus contains 1,284 manually annotated ground-truth entity instances covering nation-state group activities (APT28, APT29, APT34, Lazarus Group, MuddyWater, Sandworm, OilRig, and TA558) across 15 entity types in three functional groups. Threat actor attributes comprise `threat_actor_name` and `threat_actor_country`. Campaign context entities cover `target_country`, `target_sector`, `malware_name`, `goal`, `tool`, and `vulnerability`. Technical indicators include `ip`, `domain`, `url`, `email`, `sha256`, `sha1`, and `md5`.

Entity annotation was performed by two annotators working in the CTI function at a Computer Emergency Response Team, each independently processing all reports across all 15 entity types. Divergences were resolved through discussion until consensus was reached, following the dual-annotator reconciliation practice established by Della Penna et al. [39]. Where recurring disagreements revealed underspecified boundary conditions (most frequently at the tool and malware boundary), explicit clarification rules were formulated, documented, and applied retroactively. Controlled vocabulary was introduced for target sectors (18 STIX-aligned types) and goals (5 values) to address surface-form variability. An explicit CVE-YYYY-NNNNN format constraint was added after all three models initially returned natural-language descriptions rather than identifier strings, achieving F1 = 1.000 across all models.

The annotated corpus and the 17-actor attribution dataset used in this study can be made available to researchers upon reasonable request.

3.2 Proposed Framework

The framework comprises four sequential stages illustrated in Figure 1: (A) LLM-based NER extraction, (B) STIX 2.1 knowledge representation, (C) multi-report aggregation with Campaign Frequency scoring, and (D) CF-IDF-weighted Bayesian scoring with min-max normalization. The design directly addresses the three deficiencies identified in Section 2.1: STIX 2.1 provides the common schema; union-based aggregation with source-confidence filtering operationalizes reproducibility; and Bayesian scoring with min-max normalization produces calibrated, communicable uncertainty.

3.3 LLM-Based NER Layer

Three state-of-the-art LLM models were evaluated to select the best-performing model for extracting structured CTI data from unstructured reports. Claude Sonnet 4.6 (claude-sonnet-4-6), GPT-5.4 (gpt-5.4), and LLaMA 4 Scout (meta-llama/llama-4-scout-17b-16e-instruct) were tested. All models are queried with temperature and top-p values set to zero for deterministic outputs. Two prompting conditions, shown in Figures 2 and 3, are evaluated. In the zero-shot condition, the model receives only a task description and the target document. In the few-shot condition, the prompt is augmented with two fully annotated CTI report excerpts excluded from the evaluation corpus, providing at least one positive instance of each of the 15 entity types. Depending on the needs and token usage preferences, both approaches can be implemented in production-ready systems. All models return a structured JSON response with one array per entity type. Iterative prompt refinement added: (i) character-length constraints for SHA-256 (64 hex characters) and SHA-1 (40 hex characters), after initial evaluation revealed systematic confusion between hash types; (ii) the CVE-YYYY-NNNNN format requirement; and (iii) controlled vocabularies for sector and goal. These format-level interventions produced the largest per-entity F1 improvements, raising SHA-1 F1 by up to +0.442.

Table 1. Key limitations of representative attribution approaches.

Study	Artifact scope	Weighted	Prob. output	Key limitation
[17]	Malware binary	No	No	Single modality; no uncertainty; closed dataset
[18]	Network logs (AWS honeypot)	No	No	Single modality; binary output; no schema
[19]	TTPs, tools, malware	No	No	No network indicators; unstructured schema
[20]	Cobalt Strike command sequences	No	No	Single behavioral modality; binary output
[21]	TTPs, malware, IPs, domains, URLs, vulnerabilities, registries, filenames, file paths, emails	No	No	Deterministic classifier; no calibration; no schema
[23, 24]	TTPs, tools, malware, behavioral context, target features	No	No	Uniform ML weighting; binary output
[16]	TTPs	Association-rule-weighted Jaccard	No	Statistical weighting only; no class-based evidential weighting; no schema
[15]	Kill-chain events	No	Yes (Bayesian)	No standard schema; no calibration; manually constructed
[26]	Eight abstract simulated features	No	Yes (PMF)	Synthetic data only; abstract features; no real CTI artifacts; no standard schema
[30]	TTPs, IoCs, malware	No	No	Deterministic graph match; no uncertainty
[31]	IPs, domains, URLs, ASNs	No	No	Single modality; deterministic GNN; no schema
This work	Malware, CVEs, tools, sectors, countries, hashes, IPs, domains, URLs	CF-IDF and class-specific ε	Yes	Discussed in Section 5

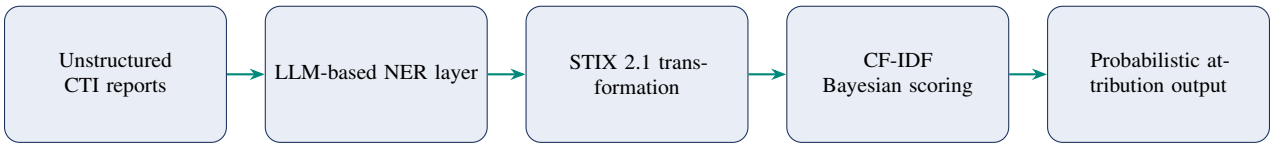


Figure 1. The integrated attribution pipeline.

You are a cybersecurity threat intelligence analyst. Your task is to extract named entities from the CTI report provided below and return only a valid JSON object that strictly conforms to the predefined schema. For any entity types not present in the document, return empty arrays ([]). Do not include any text, explanation, or formatting outside the JSON output. Ensure that all occurrences of each entity type are extracted across the entire document—do not limit extraction to the first instance. In particular, for hash-related fields, include every identified value. . .

Figure 2. Zero-shot prompt.

Excerpt: APT28, also known as Fancy Bear, is a Russian espionage group that targeted German government agencies and French energy companies. The campaign deployed X-Agent malware and leveraged CVE-2021-34527 (PrintNightmare). Mimikatz was used for credential dumping. C2 infrastructure: 185.220.101[.]45, update[.]evil-domain[.]com

Output:

```
{ "threat_actor_name": ["APT28"],
"threat_actor_country": ["russia"],
"target_country": ["germany", "france"],
"target_sector": ["government", "energy"],
"malware_name": ["x-agent"], "goal":
["Espionage"], "tool": ["mimikatz"],
"ip": ["185.220.101.45"], "domain":
["update.evil-domain.com"], "url": [], "email":
[], "sha256": [], "sha1": [], "md5": [],
"vulnerability": ["CVE-2021-34527"]}
```

Figure 3. Example given in the few-shot prompt.

3.4 STIX 2.1 Transformation Layer

Each NER output is transformed into a STIX 2.1 Bundle containing a fixed schema of nine SDO types and their associated SROs. The SDO types are Campaign, ThreatActor, Malware, Tool, Identity, Location, Vulnerability, and Indicator, where the Indicator type covers IP address, domain, URL, email, SHA-256, SHA-1, and MD5 hash. Each Indicator carries a STIX pattern derived from its type. SROs encode semantic links between these objects, specifically campaign attributed-to threat-actor, malware used-by campaign, tool used-by threat-actor, indicator indicates campaign, and threat-actor targets identity, location, and vulnerability. This fixed schema ensures that feature extraction is identical across all actors and reports, directly addressing the interoperability gap identified in Section 2.1. A visual representation of the SDO and SRO relations is shown in Figure 4.

3.5 Multi-Report Aggregation and Campaign Frequency Scoring

The aggregation layer collapses per-report feature vectors into a single actor profile using a set union:

$$F_a^{\text{agg}} = \bigcup_{r \in R(a)} F_r. \tag{1}$$

where $R(a)$ is the set of reports for actor a passing the source-confidence filter. This design choice addresses a known failure mode: treating each report as a separate known actor splits probability mass and degrades attribution performance [3]. Union aggregation is preferred over intersection, which penalizes well-documented actors with diverse campaign observables. To capture behavioral repetition that naive union aggregation destroys, a Campaign Frequency (CF) score is computed for each feature value:

$$\text{CF}_{a,f,v} = \frac{\left|\{r \in R(a) : v \in F_{r,f}\}\right|}{|R(a)|}. \tag{2}$$

A CF value of 1.0 indicates the observable appeared in every known campaign—the maximum behavioral consistency signal. This operationalizes the insight from Skopik and Pahi [7] that artifact durability correlates with attribution evidential value.

3.6 CF-IDF-Weighted Jaccard Similarity and Bayesian Scoring

Not all artifact values are equally discriminative. Observables common across many actors provide weaker attribution signal than values unique to specific actors. This observation, formally established by Hwang and Kim [27] through expert elicitation, is operationalized here through Inverse Document Frequency weighting computed over aggregated actor profiles:

$$\text{IDF}_{f,v} = \log\left(\frac{N+1}{df(f,v)+1}\right) + 1. \tag{3}$$

where N is the number of known actors and $df(f,v)$ is the count of actors whose profile contains value v for feature f. The CF-IDF-weighted Jaccard similarity integrates global rarity (IDF) and behavioral consistency (CF):

$$J_{\text{CF-IDF}}(U_f, A_{a,f}) = \frac{\sum_{v \in U_f \cap A_{a,f}} \text{IDF}_{f,v}\,\text{CF}_{a,f,v}}{\sum_{v \in U_f \cup A_{a,f}} \text{IDF}_{f,v}\,\max(\text{CF}_{a,f,v}, 1)}. \tag{4}$$

The log-posterior score for actor a_i under a uniform prior is:

$$\log S(a_i) = \sum_f w_f \log\left(J_{\text{CF-IDF}}(U_f, A_{a_i,f}) + \varepsilon_f\right). \tag{5}$$

where w_f is the base feature weight (Table 2) and ε_f is the feature-class-specific Laplace smoothing floor (Table 3). Smoothing floors directly encode the artifact perishability insight from [7, 27]: network indicators receive high floors (0.85–0.95) because their absence is uninformative since they are rotated frequently, while behavioral features receive low floors (0.05) because their absence is meaningful. This class-specific treatment distinguishes the scoring framework from the uniform-weight approaches prevalent in the literature.

Table 2. Feature weights.

Feature	Weight	Rationale
Malware	3.0	Highly actor-specific; durable behavioral fingerprint
CVEs	2.5	Rare exploitation patterns; actor capability signal
Tools	2.0	Behavioral fingerprint; offensive toolchain identity
Target Sectors	2.0	Targeting intent; stable mandate signal
File Hashes	2.0	Binary-level evidence; semi-durable
IP Addresses	1.5	Infrastructure overlap; perishable
Domains	1.5	Infrastructure overlap; perishable
Target Countries	1.5	Geographic targeting intent
Actor Country	1.0	Origin corroboration; contextual
URLs	1.0	Very perishable; minimal weight

3.7 Min-Max Normalization and Saturation Detection

The natural approach for converting log-scores into a probability distribution over candidate actors is softmax normalization. However, when one actor scores substantially above all others on a high-weight feature, softmax produces near-certain probability assignment to that actor regardless of the actual score differences—a mathematical amplification artifact, not an evidential finding. To preserve the relative ordering of candidates without inflating the leading score, min-max normalization is applied instead:

$$S_i' = \frac{\log S(a_i) - S_{\min}}{S_{\max} - S_{\min}}. \tag{6}$$

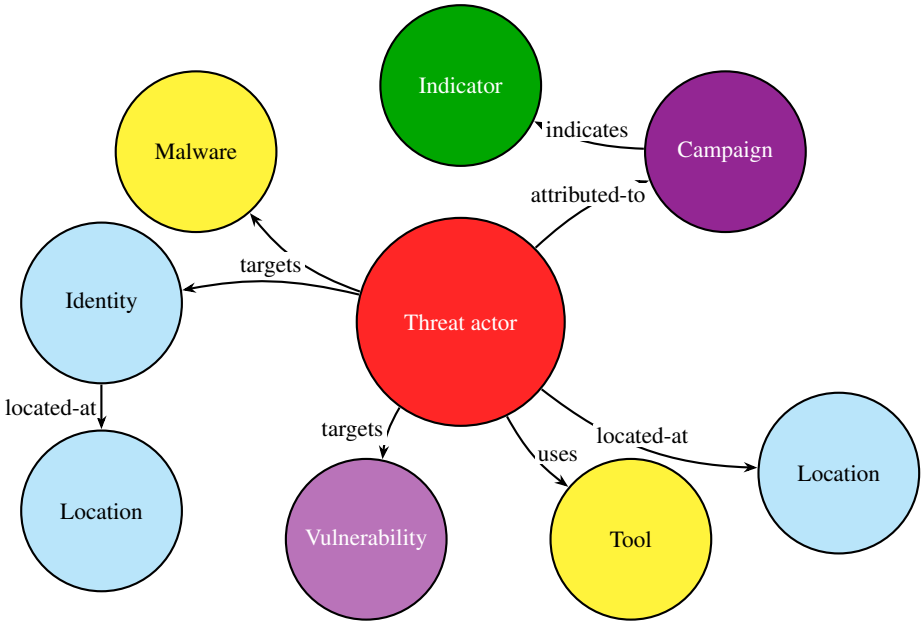


Figure 4. SRO and SDO relations.

Table 3. Feature-class-specific smoothing floors.

Feature class	ε_f	Rationale
Malware, tools, sectors, CVEs, countries	0.05	Behavioral; absence is informative
File Hashes	0.30	Semi-stable; moderate absence penalty
IP Addresses	0.90	Perishable; near-zero absence penalty
Domains	0.85	Perishable; near-zero absence penalty
URLs	0.95	Very perishable; minimal absence penalty

where $S_{\min}$ and $S_{\max}$ are the minimum and maximum log-scores across all known actors. Normalized scores are then re-scaled to sum to unity to yield relative probabilities:

$$P(a_i \mid E) = \frac{S'_i}{\sum_j S'_j}. \tag{7}$$

This preserves rank ordering, distributes probability mass across the full actor set, and does not amplify small score differences. When $S_{\max} = S_{\min}$ (all actors score identically), scores are set to $1/N$. A saturation detector flags outputs for mandatory analyst review when $P(a_1 \mid E) > 0.85$ or the gap $\Delta = P(a_1 \mid E) - P(a_2 \mid E) > 0.60$, signalling that a single feature is likely dominating the posterior.

4. EXPERIMENTAL RESULTS

4.1 NER Layer Performance

Tables 4 and 5 present macro and micro performance across all three models under both prompting conditions. Claude Sonnet 4.6 and GPT-5.4 are closely matched under zero-shot conditions, achieving micro F1 of 0.790 and 0.792, respectively. GPT-5.4 holds a slight advantage in macro F1 (0.523 vs. 0.514), reflecting marginally better performance on low-frequency entity types. Claude achieves higher micro recall (0.865 vs. 0.850), suggesting a preference for breadth of extraction. LLaMA 4 Scout trails substantially (micro F1 = 0.687), driven primarily by poor tool entity precision (0.288) and aggressive over-extraction.

Table 4. Macro NER performance.

Model and condition	Macro P	Macro R
Claude 4.6 Zero-Shot	0.516	0.533
Claude 4.6 Few-Shot	0.520	0.524
GPT-5.4 Zero-Shot	0.531	0.527
GPT-5.4 Few-Shot	0.530	0.498
LLaMA 4 Scout Zero-Shot	0.459	0.474
LLaMA 4 Scout Few-Shot	0.422	0.404

Table 5. Micro NER performance.

Model and condition	Micro P	Micro R
Claude 4.6 Zero-Shot	0.782	0.890
Claude 4.6 Few-Shot	0.802	0.875
GPT-5.4 Zero-Shot	0.799	0.866
GPT-5.4 Few-Shot	0.815	0.792
LLaMA 4 Scout Zero-Shot	0.700	0.754
LLaMA 4 Scout Few-Shot	0.746	0.548

Performance across the 15 entity types is highly non-uniform. CVE extraction achieves F1 = 1.000 across all models following the explicit format constraint, while IP address extraction reaches 0.903–0.982 F1. The tool entity type is the most challenging across all models (Claude: 0.611, GPT-5.4: 0.700, LLaMA: 0.468 under zero-shot), driven by malware and tool boundary ambiguity, system utility over-extraction (Living Off the Land Binaries), and name normalization failures that exact-match evaluation does not credit. Few-shot prompting

reduces hallucination rates by 0.028–0.037 across all models. However, GPT-5.4 exhibits a severe cardinality anchoring failure on the DynoWiper report: conditioned on exemplars containing 1–3 malware entities, it extracted exactly one malware name from a report cataloguing 22, causing micro F1 to collapse from 0.921 to 0.516. This establishes a practical design requirement: few-shot exemplars must be matched not only in entity-type coverage but also in entity-count distribution.

4.2 Attribution Results

Table 6 presents the full attribution ranking for the unknown campaign observed in public malware repositories targeting Azerbaijani and Turkish organizations, evaluated against the corpus. No saturation condition was triggered ($\Delta = 0.203$), indicating that probability mass is meaningfully distributed across candidate actors rather than collapsed onto a single winner.

Table 6. Attribution ranking.

Rank	Actor	Log score	Probability
1	TA558	−31.55	31.99%
2	MuddyWater	−36.16	11.70%
3	Sandworm	−36.99	8.04%
4	OilRig	−37.04	7.79%
5	APT34	−37.14	7.37%

TA558's top attribution rests on two high-quality matches shown in Table 7. The malware family match carries IDF = 3.197 and CF = 1.000, indicating Remcos RAT is rare across the 17-actor dataset and appeared in every documented TA558 campaign, the maximum behavioral consistency signal under the CF metric. The CVE-2017-11882 match carries IDF = 3.197 and CF = 0.333, indicating a rare exploitation pattern observed in only one of the reports covering TA558 activity. MuddyWater's case rests entirely on geographic alignment: Azerbaijan (IDF = 3.197, CF = 0.400) and Turkey (IDF = 2.792, CF = 0.600) fall within MuddyWater's established South Caucasus and MENA operational footprint. The combined geographic signal is outweighed by TA558's high-IDF, high-CF malware match under the base weight assignments ($w_{\text{malware}} = 3.0$ and $w_{\text{target_countries}} = 1.5$). TA558's lack of geographic overlap with Azerbaijan or Turkey in its three public reports is labeled in the pipeline output as potentially reflecting reporting gaps rather than behavioral absence, consistent with the dark figure advisory mechanism.

Table 7. Contributing features for top attribution candidates.

Actor	Feature	Value	IDF
TA558	Malware	Remcos RAT	3.197
TA558	CVE	CVE-2017-11882	3.197
TA558	Sector	Financial	1.693
MuddyWater	Country	Azerbaijan	3.197
MuddyWater	Country	Turkey	2.792
MuddyWater	Sector	Transportation	2.504
MuddyWater	Sector	Financial	1.693

5. DISCUSSION

The pipeline's primary operational value is to support, not replace, human attribution analysis. Full score transparency, with every scoring decision traceable to specific feature matches and explicit IDF and CF values, enables an analyst with classified knowledge to rationally override the geographic divergence signal. This transparency property, absent from the black-box deep learning approaches reviewed in Section 2.3, is the most operationally significant characteristic of the framework.

The NER evaluation establishes that entity types are a stronger predictor of extraction difficulty than model identity: the rank ordering of entity types by F1 is largely preserved across all three architecturally distinct models. This confirms a domain-level challenge, not a model-level one. Entities with machine-readable formats are amenable to near-perfect extraction through explicit format specification alone. Entities requiring operational domain knowledge to disambiguate—tool versus malware, implied sector versus named sector—resist improvement through either few-shot conditioning or format constraints, pointing toward fine-tuning on annotated CTI corpora such as CyberNER [37] or CTI-HAL [39] as the necessary next step.

The framework is not without limitations. The Naive Bayes independence assumption treats all features as statistically uncorrelated; in practice, geographic targeting, sector targeting, and actor origin are structurally correlated for nation-state actors operating under geopolitically defined mandates [13]. A Bayesian network model would capture these dependencies but requires substantially more training data than is available in open-source corpora [3]. The IDF weights are computed over the 17-actor dataset; as the corpus grows, weights will converge toward their true discriminative values. The pipeline operates exclusively on public CTI, so the true behavioral profile of any actor extends beyond what vendor reports document—the dark figure problem, demonstrated by Pahi and Skopik [2] to be a fundamental limit on technical attribution, is irreducible within any open-source attribution system.

Further structural limitations concern threat actor naming conventions. Saha et al. report that inconsistent names impede merging and correlating threat information from disparate sources [8]. Without alias resolution, reports about the same actor published under different vendor-assigned names are treated as distinct profiles, fragmenting the evidence base and inflating the apparent actor count. STIX 2.1 offers a natural mechanism for addressing this through the aliases property of the ThreatActor SDO, which allows multiple vendor-assigned names to be declared equivalent within a single object. Incorporating alias resolution before aggregation, through a curated mapping or automated entity linking, would consolidate fragmented profiles and improve IDF calibration and attribution accuracy.

6. CONCLUSION

This paper presented an integrated end-to-end cyber threat attribution pipeline combining LLM-based named entity recognition, STIX 2.1 knowledge representation, and Bayesian probabilistic scoring. The framework directly addresses

the three deficiencies identified in the literature: non-reproducibility is resolved through STIX 2.1 standardization and union-based multi-report aggregation; the schema gap is addressed by grounding all feature extraction in a fixed nine-object STIX bundle; and probabilistic uncertainty is communicated through min-max normalized Bayesian scores with an explicit dark figure advisory. The NER evaluation demonstrates that commercial LLMs achieve zero-shot micro F1 of approximately 0.80 without task-specific fine-tuning. A single prompt revision—adding explicit character-length constraints for hash entities—produced the largest per-entity F1 improvements, establishing format specification as the highest-leverage intervention for technically structured entity types.

The attribution engine operationalizes a contribution absent from all reviewed systems: differential artifact weighting across feature classes, grounded in expert consensus from the attribution literature, through Campaign Frequency scoring, IDF-adjusted rarity weighting, and class-specific Laplace smoothing floors that treat behavioral and perishable artifacts differently. Case-study evaluation against 17 actors and 57 reports produced probability distributions, surfacing an analytical ambiguity between toolchain evidence (favoring TA558) and geographic behavioral evidence (favoring MuddyWater) that binary attribution methods would have suppressed.

Future work will address the Naive Bayes independence assumption through Bayesian network modeling, incorporate temporal decay for perishable indicators, develop a confidence calibration layer mapping scores to empirically validated posterior probabilities, and evaluate the pipeline on a larger actor corpus spanning additional regional threat groups.

REFERENCES

[1] T. Rid and B. Buchanan. Attributing cyber attacks. *Journal of Strategic Studies*, 38(1–2):4–37, 2015.

[2] T. Pahi and F. Skopik. Cyber attribution 2.0: Capture the false flag. In *Proceedings of the 18th European Conference on Cyber Warfare and Security (ECCWS 2019)*, pages 338–345, 2019.

[3] N. Rani, B. Saha, and S. K. Shukla. A comprehensive survey of automated advanced persistent threat attribution: Taxonomy, methods, challenges and open research problems. *Journal of Information Security and Applications*, 92:104076, 2025.

[4] H. Alturkistani, A. G. Jaafar, and S. Chuprat. Automating cyber threat intelligence workflows with LLMs: Processing, analyzing, and defending in one model. In *2025 3rd International Conference on Cyber Resilience (ICCR)*, pages 1–6, Dubai, United Arab Emirates, 2025.

[5] K. Guru, R. J. Moss, and M. J. Kochenderfer. On technique identification and threat-actor attribution using LLMs and embedding models. In *2025 International Conference on Cybersecurity and AI-Based Systems (Cyber-AI)*, pages 332–339, Varna, Bulgaria, 2025.

[6] V. Mavroeidis and S. Bromander. Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 91–98, Athens, 2017.

[7] F. Skopik and T. Pahi. Under false flag: Using technical artifacts for cyber attack attribution. *Cybersecurity*, 3(1):8, 2020.

[8] A. Saha, J. Mattei, J. Blasco, L. Cavallaro, D. Votipka, and M. Lindorfer. Expert insights into advanced persistent threats: Analysis, attribution, and challenges. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2185–2204, Seattle, WA, 2025.

[9] D. Blänsdorf. Threat attribution using generative AI: A generative AI-based tool to support the threat attribution process. Master's thesis, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden, 2024.

[10] N. Prasad, A. Diro, M. Warren, and M. Fernando. A survey of cyber threat attribution: Challenges, techniques, and future directions. *Computers & Security*, 157:104606, 2025.

[11] P. Deka, S. Rajapaksha, R. Rani, A. Almutairi, and E. Karafili. AttackER: Towards enhancing cyber-attack attribution with a named entity recognition dataset. arXiv:2408.05149, 2024.

[12] S. Caltagirone, A. Pendergast, and C. Betz. The diamond model of intrusion analysis, 2013.

[13] P. N. Bahrami, A. Dehghantanha, T. Dargahi, R. M. Parizi, K.-K. R. Choo, and H. H. S. Javadi. Cyber kill chain-based taxonomy of advanced persistent threat actors: Analogy of tactics, techniques, and procedures. *Journal of Information Processing Systems*, 15(4):865–889, 2019.

[14] Q. Li, Z. Yang, Z. Jiang, B. Liu, and Y. Fu. Association analysis of cyber-attack attribution based on threat intelligence. In *Proceedings of the 2nd Joint International Information Technology, Mechanical and Electronic Engineering Conference (JIMEC 2017)*, volume 62 of *Advances in Computer Science Research*, pages 222–230. Atlantis Press, 2017.

[15] Q. Li, Z.-M. Yang, B.-X. Liu, and Z.-W. Jiang. A reasoning method of cyber-attack attribution based on threat intelligence. *International Journal of Computer and Systems Engineering*, 10(5):920–924, 2016.

[16] K. Edie, C. Mckee, and A. Duby. Extending threat playbooks for cyber threat intelligence: A novel approach for APT attribution. In *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6, Chattanooga, TN, USA, 2023.

[17] I. Rosenberg, G. Sicard, and E. David. DeepAPT: Nation-state APT attribution using end-to-end deep neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2017*, volume 10614 of *Lecture Notes in Computer Science*, pages 91–99, Alghero, Italy, 2017.

[18] M. Alkaabi and S. O. Olatunji. Modeling cyber-attribution using machine learning techniques. In *2020 30th International Conference on Computer Theory and Applications (ICCTA)*, pages 10–15, Alexandria, Egypt, 2020.

[19] U. Noor, Z. Anwar, T. Amjad, and K.-K. R. Choo. A machine learning-based fintech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems*, 96:227–242, 2019.

[20] E. Böge, M. B. Ertan, H. Alptekin, and O. Çetin. Unveiling cyber threat actors: A hybrid deep learning approach for behavior-based attribution. *Digital Threats*, 6(1):1–20, 2025.

[21] N. Xiao, B. Lang, T. Wang, and Y. Chen. APT-MMF: An advanced persistent threat actor attribution method based on multimodal and multilevel feature fusion. *Computers & Security*, 144:103960, 2024.

[22] X. Xiang, H. Liu, L. Zeng, H. Zhang, and Z. Gu. IPAttributor: Cyber attacker attribution with threat intelligence-enriched intrusion data. *Mathematics*, 12(9):1364, 2024.

[23] E. Irshad and A. B. Siddiqui. Cyber threat attribution using unstructured reports in cyber threat intelligence. *Egyptian Informatics Journal*, 24(1):43–59, 2023.

[24] E. Irshad and A. B. Siddiqui. Context-aware cyber-threat attribution based on hybrid features. *ICT Express*, 10(3):553–569, 2024.

[25] M. R. Ali Khan and A. Almulhem. Behavioral and propagation-based analysis of APT attacks for effective attack attribution. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–8, Boston, MA, USA, 2025.

[26] K. T. W. Teuwen. A modular approach to automatic cyber threat attribution using opinion pools. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3089–3098, Sorrento, Italy, 2023.

[27] S. Hwang and T.-S. Kim. An exploratory study on artifacts for cyber attack attribution considering false flag: Using delphi and AHP methods. *IEEE Access*, 11:74533–74544, 2023.

[28] S. Bromander et al. Investigating sharing of cyber threat intelligence and proposing a new data model for enabling automation in knowledge representation and exchange. *Digital Threats*, 3(1):1–22, 2022.

[29] F. Marchiori, M. Conti, and N. V. Verde. STIXnet: A novel and modular solution for extracting all STIX objects in cti reports. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–11, Benevento, Italy, 2023.

[30] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian. CSKG4APT: A cybersecurity knowledge graph for advanced persistent threat organization attribution. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5695–5709, 2023.

[31] I. J. King, R. Ramirez, B. Bowman, and H. H. Huang. TRAIL: A knowledge graph-based approach for attributing advanced persistent threats. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 1207–1220, Hong Kong, Hong Kong, 2025.

[32] P. Evangelatos et al. Named entity recognition in cyber threat intelligence using transformer-based models. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 348–353, Rhodes, Greece, 2021.

[33] S. Zhou, J. Liu, X. Zhong, and W. Zhao. Named entity recognition using BERT with whole word masking in cybersecurity domain. In *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*, pages 316–320, Xiamen, China, 2021.

[34] S.-S. Chen, R.-H. Hwang, C.-Y. Sun, Y.-D. Lin, and T.-W. Pai. Enhancing cyber threat intelligence with named entity recognition using BERT-CRF. In *GLOBECOM 2023: 2023 IEEE Global Communications Conference*, pages 7532–7537, Kuala Lumpur, Malaysia, 2023.

[35] P. Ranade, A. Piplai, A. Joshi, and T. Finin. CyBERT: Contextualized embeddings for the cybersecurity domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3334–3342, Orlando, FL, USA, 2021.

[36] E. Aghaei, S. Jain, P. Arun, and A. Sambamoorthy. SecureBERT 2.0: Advanced language model for cybersecurity intelligence. In *2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops)*, pages 382–388, Honolulu, HI, USA, 2025.

[37] Y. Ech-Chammakhy, A. Motii, A. Rabii, O. Azrara, and J. Chbili. CyberNER: A harmonized STIX corpus for cybersecurity named entity recognition. In *2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 2190–2197, Guiyang, China, 2025.

[38] H.-J. Chan, C.-Y. Hsu, C.-C. Chien, J.-J. Wu, and H.-L. Ku. FeedRef2022: A named entity recognition dataset for extracting indicators of compromise. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2578–2584, Osaka, Japan, 2022.

[39] S. Della Penna, R. Natella, V. Orbinato, L. Parracino, and L. Pianese. CTI-HAL: A human-annotated dataset for cyber threat intelligence analysis. In *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 69–78, Venice, Italy, 2025.

[40] D. Plohmann, M. Clauß, S. Enders, and E. Padilla. Malpedia: A collaborative effort to inventorize the malware landscape. *The Journal on Cybercrime and Digital Investigations*, 3(1):1–19, 2017.