

SMAH-CTI: A Unified Sentence-Level Benchmark for Multi-Task Cyber Threat Intelligence Extraction

Niveen O. Jaffal¹, Radi Jarrar^{1,3}, Mohammed Alkhanafseh¹, and David Mohaisen^{2,3}

¹Birzeit University

²University of Central Florida

³Corresponding Author

Abstract

Cyber Threat Intelligence (CTI) reports describe adversary behavior and attack infrastructure in natural language but remain largely unstructured and difficult to process at scale. Existing CTI benchmarks are limited in realism or scope, often relying on synthetic or knowledge-base-derived data, coarse document-level annotations, or isolated subtasks. We introduce SMAH-CTI, a human-annotated benchmark derived from 100 public CTI reports from five threat intelligence providers. It contains 12,304 validated sentence-level instances annotated for (i) relevance classification, (ii) CTI-specific named entity recognition (NER), and (iii) multi-label MITRE ATT&CK tactic, technique, and sub-technique mapping, together with grounded procedure descriptions. By unifying these tasks over evidence grounded in real CTI reports, SMAH-CTI provides a reproducible benchmark for evaluating structured CTI extraction and transforming unstructured reports into machine-interpretable intelligence.

1 Introduction

Cyber Threat Intelligence (CTI) refers to the systematic collection and analysis of information about current and emerging threats to support proactive defense. A substantial threat intelligence is disseminated in unstructured natural-language formats, including vendor reports, incident response documentation, security advisories, technical blogs, and open-source artifacts. Because these sources are unstructured, analyzing them at scale remains difficult (Santos et al., 2025).

CTI characterizes adversarial behavior through Tactics, Techniques, and Procedures (TTPs), providing richer behavioral insight than volatile indicators of compromise (IoCs). MITRE ATT&CK (MITRE Corporation, 2025) provides a standard taxonomy for representing these behaviors as tactics, techniques, and sub-techniques.

Mapping CTI narratives to ATT&CK identifiers enables semantic normalization across heterogeneous reports and systematic threat analysis. However, this mapping remains challenging due to diverse writing styles, implicit descriptions of adversarial behavior, and the need for fine-grained semantic interpretation (Rahman et al., 2025).

Problem Statement. Despite growing interest in automated CTI extraction, existing benchmarks remain limited in realism, granularity, and task coverage. Many rely on knowledge bases or synthetic text rather than analyst-authored reports, focus on isolated tasks, or use coarse document-level annotations. Even real-world datasets often lack sentence-level grounding and unified multi-label supervision across ATT&CK tactics, techniques, and sub-techniques. These limitations hinder reproducible and systematic evaluation, particularly under the linguistic variability, implicit behaviors, long-tail labels, and cross-task dependencies characteristic of real CTI reports.

To address these limitations, we introduce SMAH-CTI, a human-annotated sentence-level benchmark derived from real-world CTI reports. The dataset supports unified multi-task supervision over relevance classification, CTI-specific named entity recognition (NER), and multi-label ATT&CK mapping, together with grounded procedure descriptions linked to source sentences. By combining sentence-level grounding with authentic CTI data, SMAH-CTI enables reproducible evaluation of grounded multi-task CTI extraction under realistic reporting conditions.

Contributions. We contribute: (1) A *human-annotated, sentence-level CTI benchmark* comprising 12,304 validated instances from diverse real-world CTI reports; (2) A *unified, grounded multi-task annotation framework* covering relevance, CTI entities, ATT&CK tactics, techniques, sub-techniques, and evidence-grounded procedure descriptions; (3) A *reproducible evaluation pro-*

*to*col with deterministic identifiers, provenance metadata, fixed report-level splits, and standardized release formats; and (4) *Baseline evaluation and diagnostic analysis* spanning single and multi task, and instruction-following models, with analyses of relevance gating, multi-label complexity, long-tail unseen labels, and representative errors.

2 Background

This section provides background on cyber threat intelligence (CTI), CTI extraction tasks, and the MITRE ATT&CK framework.

CTI and Security Tasks. CTI is evidence-based knowledge about existing and emerging cyber threats used to support security operations and decision making (Kotsias et al., 2023). To facilitate sharing and automation, several structured CTI standards have been developed, including STIX (OASIS, 2021a), TAXII (OASIS, 2021b), and MITRE ATT&CK (MITRE Corporation, 2025). However, much operational CTI remains embedded in unstructured natural-language reports produced by security vendors, incident response teams, and open-source intelligence communities, making it difficult to search, normalize, and operationalize at scale.

Traditional security analysis has largely focused on malware execution traces rather than narrative CTI. For example, behavior-centric frameworks such as AMAL (Mohaisen et al., 2015) characterize malware from runtime artifacts, including file-system, registry, memory, and network activity, but do not directly extract structured adversarial knowledge from natural-language reports.

Automated CTI processing involves several security-oriented NLP tasks. Relevance classification distinguishes operationally relevant adversarial behavior from contextual content (Kuehn, 2025). NER extracts CTI-specific entities such as threat actors, malware, infrastructure, and adversarial actions, while relation extraction identifies semantic links among them (Marinho and Holanda, 2023). Another task is TTP extraction and mapping, which aligns narrative CTI with standard adversarial behaviors in frameworks such as MITRE ATT&CK (Ali and Peng, 2024). Other tasks include CTI summarization and structured threat knowledge-graph construction (Peng et al., 2026). Together, these tasks motivate the unified multi-task framework adopted in SMAH-CTI.

MITRE ATT&CK Framework. MITRE

ATT&CK is a publicly available knowledge base that organizes adversarial behavior into tactics, techniques, and sub-techniques derived from real-world threat observations (MITRE Corporation, 2025). Tactics represent higher-level adversarial objectives, while techniques describe behaviors used to achieve those objectives. Sub-techniques capture more specific operational variants; e.g., T1136 (Create Account) is a general technique, whereas T1136.001 (Local Account) specifies a particular variant (MITRE Corporation, 2025).

MITRE maintains ATT&CK matrices for Enterprise, Mobile, and Industrial Control Systems (ICS). We focus on Enterprise ATT&CK because it is most relevant to the enterprise-oriented CTI reports considered in this work. Enterprise ATT&CK v18.1 contains 14 tactics, 216 techniques, and 475 sub-techniques (MITRE Corporation, 2025) and serves as the reference ontology for SMAH-CTI annotation and evaluation.

3 Related Work

Table 1 lists existing datasets along these dimensions and contrasts them with SMAH-CTI.

3.1 CTI Extraction Benchmarks

Existing CTI benchmarks vary substantially in data source, annotation granularity, and task formulation. Several rely on structured knowledge bases, curated ATT&CK descriptions, or synthetic text, not analyst-authored reports. CTI-to-MITRE (Orbinato et al., 2022), TTPHunter (Rani et al., 2023), and TTPXHunter (Rani et al., 2024), for example, primarily address ATT&CK technique classification using curated or synthetically augmented data, without unified supervision for relevance, NER, and grounded procedures. TIEF (Joy et al., 2025) similarly targets multi-label TTP classification using ATT&CK-aligned data augmented with LLM-generated samples.

Others draw from real-world CTI reports but differ in annotation granularity and scope. rcATT (Legoy et al., 2020) supports multi-label report-level ATT&CK classification without sentence grounding, while aCTIon (Siracusano et al., 2023) provides STIX annotations but remains document-centric. At finer granularity, CTI-HAL (Della Penna et al., 2025) provides sentence-level multi-label ATT&CK annotations, whereas LADDER (Alam et al., 2023) provides entity- and relation-oriented extraction for attack-pattern

Table 1: Comparison of representative CTI extraction benchmarks and SMAH-CTI across data source, annotation granularity, task scope, and ATT&CK-aligned supervision.

Benchmark	Year	Human Ann.	Source			Granularity					Unified	Task Scope			ATT&CK Supervision			
			R	KB	Mi	Sp	Se	Wi	Do	Ot		Re	NER	Pr	Ta	Te	St	ML
rcATT (Legoy et al., 2020)	2020		•						•						✓	✓		✓
Cti-To-Mitre (Doc.) (Orbinato et al., 2022)	2022				•				•							✓		✓
Cti-To-Mitre (Sent.) (Orbinato et al., 2022)	2022			•			•									✓		
TIM (You et al., 2022)	2022	✓	•					•							✓	✓		
aCTIon (Siracusano et al., 2023)	2023	✓	•						•			✓			✓	✓		✓
LADDER (Alam et al., 2023)	2023	✓	•			•						✓			✓	✓		
SMET (Abdeen et al., 2023)	2023	△			•		•								✓	✓		✓
TTPHunter (Doc.) (Rani et al., 2023)	2023		•						•							✓		✓
TTPHunter (Sent.) (Rani et al., 2023)	2023			•			•									✓		✓
CTI-ATE (Alam et al., 2024)	2024	✓			•				•						✓	✓		✓
TTPXHunter (Doc.) (Rani et al., 2024)	2024	✓	•						•			✓			✓	✓		✓
TTPXHunter (Sent.) (Rani et al., 2024)	2024			•			•								✓	✓		
CTI-HAL (Della Penna et al., 2025)	2025	✓	•				•								✓	✓	✓	✓
CTI-TE (Zhang et al., 2025)	2025		•							•					✓	✓	✓	✓
Re-CTI (Zhang et al., 2025)	2025		•							•					✓	✓	✓	✓
TIEF (Joy et al., 2025)	2025			•			•							✓	✓	✓	✓	✓
CTI-Thinker (Yang et al., 2026)	2026	△	•				•					✓			✓	✓		
SMAH-CTI	2026	✓	•				•				✓	✓	✓	✓	✓	✓	✓	✓

Abbreviations: R = reports; KB = knowledge bases; Mi = mixed; Sp = span; Se = sentence; Wi = window; Do = document; Ot = other; Re = relevance classification; NER = named-entity recognition; Pr = provenance/grounded extraction; Ta = tactic; Te = technique; St = sub-technique; ML = multi-label.

knowledge construction.

Several resources address narrower CTI extraction settings. TIM (You et al., 2022) uses coarse TTP categories, while CTI-ATE (Alam et al., 2024) focuses primarily on ATT&CK technique extraction. CTI-Thinker (Yang et al., 2026), SMET (Abdeen et al., 2023), and Re-CTI (Zhang et al., 2025) provide additional CTI extraction or ATT&CK-oriented supervision, but none combines human-annotated real-world reports with sentence-level relevance, NER, grounded procedures, and fine-grained multi-label ATT&CK supervision in a unified benchmark.

3.2 Benchmark Limitations

Across these benchmarks, three limitations recur. First, knowledge-base-derived, curated, or synthetic datasets may not capture the linguistic variability and implicit behavior of analyst-authored CTI reports. Second, document-level annotations weaken the correspondence between structured outputs and the textual evidence supporting them. Third, most benchmarks target individual extraction tasks rather than jointly representing relevance, entities, procedures, and ATT&CK behaviors under a consistent annotation schema.

SMAH-CTI addresses these limitations through human-annotated, sentence-level instances derived from real-world CTI reports. It integrates relevance classification, CTI-specific NER, grounded procedure descriptions, and multi-label ATT&CK tactic, technique, and sub-technique

mapping within a unified annotation framework.

3.3 Why Sentence-Level Annotation

Annotation granularity determines how precisely adversarial behavior can be represented and grounded in source text. Document-level annotations may aggregate content, weakening alignment between structured outputs and supporting evidence, whereas span-level annotations may not capture complete behavioral semantics (Tamanna et al., 2026). Sentence-level annotation provides a balance between semantic completeness and fine-grained grounding. CTI sentences often describe localized adversarial behaviors or procedural steps, allowing relevance labels, entities, procedures, and ATT&CK mappings to be tied to a common semantic unit (Chen et al., 2023). Thus, SMAH-CTI adopts the sentence as its annotation unit for grounded multi-task CTI extraction.

4 Dataset Construction Pipeline

Figure 1 summarizes SMAH-CTI pipeline.

4.1 Data Collection

SMAH-CTI is constructed from 100 publicly available CTI reports collected between 2022 and 2026 from five threat intelligence providers: APTnotes (APTnotes, n.d.), CrowdStrike (CrowdStrike, 2024), Mandiant (Mandiant, 2024), MITRE (MITRE Corporation, 2025), and SentinelOne (SentinelOne, 2024). The corpus spans diverse behaviors, including APT campaigns,

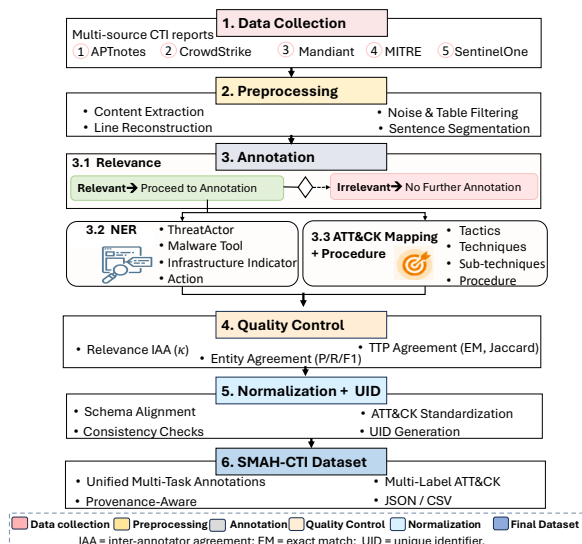


Figure 1: SMAH-CTI pipeline.

espionage operations, ransomware incidents, intrusion activity, and emerging malware trends. The multi-vendor design mitigates source-specific stylistic biases while capturing heterogeneous CTI narratives. To reflect operational reporting practices, the corpus includes both PDF and HTML reports. PDF reports preserve provenance and formatting structure but contain layout artifacts, while HTML reports provide cleaner semantic structure but lack consistent pagination.

4.2 Preprocessing Pipeline

The preprocessing pipeline converts heterogeneous CTI reports into normalized sentence-level instances while preserving semantic content, provenance, and local narrative.

Extraction and Normalization. PDF reports are processed page by page to preserve provenance, while HTML reports are parsed using structured extraction pipelines that remove non-content components. Additional preprocessing includes noise filtering, IoC-heavy content filtering, whitespace normalization, and semantic line reconstruction.

Sentence-Level Processing. We adopt sentence-level modeling because CTI reports frequently interleave operational behavior with mitigations, vulnerabilities, and descriptive narrative. Document-level annotations aggregate mixed content, while paragraph-level representations combine semantically distinct behaviors.

In contrast, sentences capture localized semantic units describing actions or ATT&CK-aligned behaviors. Sentence-level processing enables explicit grounding of relevance labels, entities, procedures, and ATT&CK mappings to a single se-

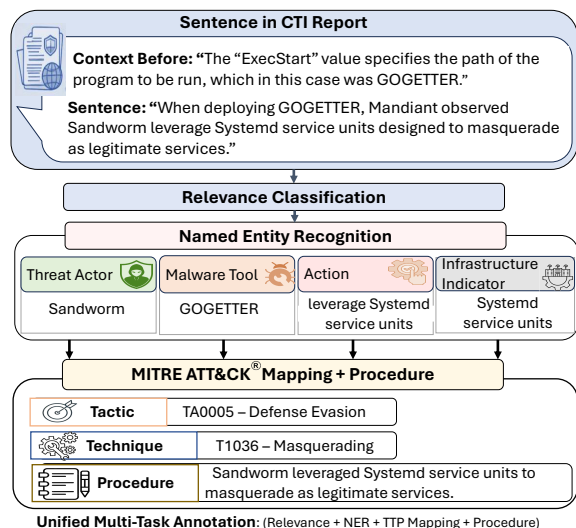


Figure 2: Example of sentence-level CTI annotation and structured MITRE ATT&CK mapping.

mantic unit while reducing ambiguity introduced by broader contexts. Sentence-level modeling also enables conditional extraction, where NER and ATT&CK mapping are applied only to sentences containing explicit adversarial behavior.

Following that, the pipeline produced 12,974 candidate sentence-level instances. Each instance preserves provenance metadata including `doc_id`, `page_number`, `sentence_id`, and `context_before`. After annotation consolidation and quality-control validation, the final SMAH-CTI release contains 12,304 validated sentence-level instances.

4.3 Annotation Schema

SMAH-CTI adopts a unified multi-task annotation schema (Figure 2) comprising: (i) relevance classification, (ii) CTI-specific named entity recognition (NER), (iii) multi-label ATT&CK mapping, and (iv) grounded procedure annotation.

Relevance Classification. Each sentence is binary-labeled indicating whether it contains operational behavior, e.g., execution, persistence, lateral movement, or command and control.

NER. Relevant sentences are annotated with four entity categories: `THREATACTOR`, `MALWARE-TOOL`, `INFRASTRUCTURE_INDICATOR`, and `ACTION`. `ACTION` captures semantically complete adversarial verb phrases rather than isolated verbs.

Multi-label TTP Mapping. Each relevant sentence is mapped to one or more ATT&CK techniques or sub-techniques.

Procedure Annotation. Each relevant sentence includes a grounded description of the instance-

specific realization of adversarial behavior.

STIX Compatibility. We adopt a compact task-oriented entity schema rather than the complete STIX object model to reduce annotation complexity. ThreatActor aligns with the STIX Threat Actor object; MalwareTool with Malware or Tool, depending on context; and Infrastructure Indicator with Infrastructure or Indicator. Action captures adversarial behavior and complements ATT&CK annotations rather than representing a standalone STIX object. This design preserves STIX interoperability while providing a lightweight schema for sentence-level benchmarking.

4.4 Manual Annotation Protocol

Annotation was conducted in Label Studio by 12 trained annotators with backgrounds in cybersecurity and NLP. Annotators followed guidelines defining relevance, entity boundaries, ATT&CK mapping, and procedure-writing conventions. The protocol is evidence-driven: ATT&CK mappings are assigned only when supported by the sentence, while speculative mappings are excluded. Uncertain cases were resolved through calibration and adjudication by a senior reviewer. Annotation proceeded hierarchically from sentence-level relevance to entities, ATT&CK mappings, and grounded procedures. Annotation time was not systematically recorded; therefore, total annotation hours cannot be reliably reported.

4.5 Inter-Annotator Agreement

Inter-annotator agreement (IAA) was evaluated on 1,000 sentences independently annotated by two annotators; detailed results and confusion matrices appear in Appendix F.4. Relevance classification achieved 89.80% agreement and Cohen’s $\kappa = 0.751$, indicating substantial agreement (Lan-dis and Koch, 1977). Technique-level ATT&CK mapping achieved 81.5% exact match and 0.816 mean Jaccard similarity, while tactic-level agreement reached 85.1% and 0.852, respectively.

4.6 Ground Truth Normalization

After annotation, outputs were consolidated into a unified schema with deterministic identifiers and provenance metadata. ATT&CK annotations were normalized to official identifiers and names, with entities represented as structured spans. Cross-layer consistency constraints required irrelevant sentences to contain no entities, ATT&CK mappings, or procedure descriptions.

Table 2: Source composition of SMAH-CTI.

Source	Rep.	Sent.	Rel.	IDs	Base	Sub.
APTnotes	21	2,106	526	101	74	36
CrowdStrike	12	2,564	670	205	115	103
Mandiant	18	2,251	727	152	96	64
MITRE	27	3,766	1,803	259	150	122
SentinelOne	22	1,617	576	165	110	65
Total	100	12,304	4,302	383	195	196

Note: Rep., Sent., and Rel. denote reports, sentences, and behaviorally relevant sentences. IDs, Base, and Sub. denote unique ATT&CK identifiers, base techniques, and sub-techniques, respectively.

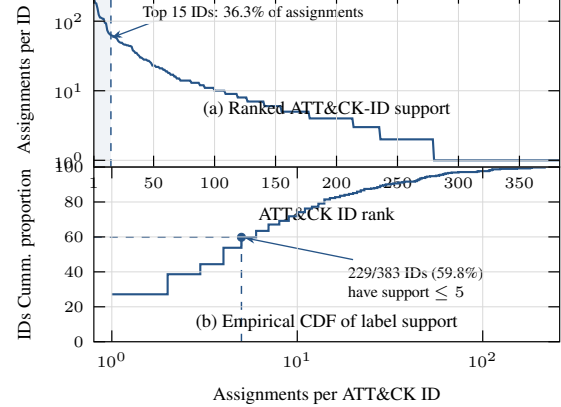


Figure 3: Long-tail distribution of ATT&CK technique and sub-technique assignments in SMAH-CTI. (a) The 15 most frequent of 383 valid ATT&CK IDs account for 36.3% of 4,837 multilabels. (b) The median support is four, and 229 IDs (59.8%) occur five times or fewer.

4.7 Dataset Statistics

SMAH-CTI has 12,304 validated sentence-level instances from 100 CTI reports, including 4,302 behaviorally relevant sentences. The dataset covers all 14 Enterprise ATT&CK tactics and exhibits a naturally imbalanced, long-tailed distribution of fine-grained techniques and sub-techniques. Table 2 summarizes the source composition of the frozen dataset across five CTI providers, revealing substantial variation in report length, relevant-sentence density, and ATT&CK coverage.

Full TTP Label Distribution. Across the dataset, gold annotations contain 4,837 sentence-label assignments spanning 383 valid ATT&CK technique and sub-technique identifiers. Each assignment links one identifier to one sentence, with multi-label sentences contributing multiple assignments. Median per-identifier support is four, and 229 identifiers (59.8%) occur five times or fewer. The 15 most frequent identifiers account for 36.3% of assignments, with the remaining 63.7% distributed across 368 identifiers. Figure 3 shows the ranked support distribution and empirical cumulative distribution of per-label support.

4.8 Experimental Protocol

We use a fixed report-level partition of SMAH-CTI to prevent leakage from shared provenance and narrative context. With master seed 42, the 100 reports are divided into 70 training, 10 development, and 20 protected test reports, yielding 8,635, 1,238, and 2,431 sentence-level instances, respectively (Table 3). No report identifier or sentence UID occurs across splits. Relevance classification and NER use all instances. Tactic prediction uses 3,006/434/862 train/development/test instances with tactic supervision, while technique- and sub-technique-level TTP mapping uses 2,902/421/841 instances with at least one valid ATT&CK identifier. The development split supports model selection and threshold tuning; the protected test split is reserved for final frozen evaluation.

Table 3: Report-level split statistics for SMAH-CTI.

Split	Reports	Sent.	Rel.	Non-rel.	Rel. (%)
Train	70	8,635	3,006	5,629	34.81
Dev	10	1,238	434	804	35.06
Test	20	2,431	862	1,569	35.46
Total	100	12,304	4,302	8,002	34.96

5 Experimental Evaluation

5.1 Experimental Setup

All baselines use the frozen report-level partitions and task-specific instance populations defined in Section 4.8. Relevance classification and NER use all sentence-level instances, with irrelevant sentences containing no gold entities. Tactic prediction uses instances with at least one tactic, while technique- and sub-technique-level ATT&CK mapping uses instances with at least one valid identifier. Model and checkpoint selection and, where applicable, decision-threshold tuning are performed exclusively on the development split. The protected test split is reserved for final frozen evaluation. All trainable single- and multi-task baselines use a single predefined run with random seed 42; we report no multi-seed averages or variance estimates. The instruction-following baseline uses fixed task-specific prompts, temperature zero, and a predefined JSON schema for consistent parsing and scoring.

5.2 Baseline Models

We evaluate four baselines spanning three modeling paradigms: single-task supervised learn-

ing, shared-encoder multi-task learning, and instruction-following LLM prompting.

Single-task RoBERTa-base. We fine-tune separate RoBERTa-base models for relevance, NER, tactic prediction, and technique- and sub-technique-level ATT&CK mapping. Relevance is formulated as binary classification, NER as BIO token classification over four entity types, and tactic and TTP prediction as multi-label classification. The TTP classifier predicts ATT&CK identifiers observed in training.

Single-task SecBERT. We fine-tune SecBERT, a cybersecurity-pretrained encoder, using the same single-task formulations and partitions as RoBERTa-base. This comparison isolates the effect of domain-specific pretraining while holding the experimental protocol fixed.

Shared-encoder multi-task RoBERTa. We evaluate a multi-task model with a shared RoBERTa-base encoder and task-specific heads for relevance, NER, tactic prediction, and TTP mapping, jointly optimized using their respective objectives.

GPT-4o-mini prompting. We evaluate GPT-4o-mini without fine-tuning using fixed task-specific prompts and temperature zero. Predictions follow a predefined JSON schema for deterministic parsing and scoring.

These baselines span complementary NLP paradigms and provide reproducible reference points across learning settings. SMAH-CTI enables standardized evaluation of specialized models under a unified protocol rather than comparison with all CTI-specific extraction systems.

5.3 Evaluation Metrics

For all tasks, precision, recall, and F1-score are computed as $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$, and $F1 = \frac{2PR}{P+R}$. For relevance classification, we report F1-score for the relevant class. For NER, we report exact span-level precision, recall, and F1-score; a prediction is correct only if its entity type and character offsets match the gold annotation.

Tactic prediction and technique- and sub-technique-level ATT&CK mapping are evaluated using micro-F1 and macro-F1. Micro-F1 aggregates counts across labels, while macro-F1 averages label-wise F1-scores, $\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} F1_{\ell}$, where $\mathcal{L}$ is the ATT&CK label set. Tactic prediction covers all 14 Enterprise ATT&CK tactics. Technique- and sub-technique-level predictions use normalized ATT&CK identifiers and are evaluated over

the complete 383-identifier space and separately over train-seen and train-unseen identifiers.

5.4 Results and Discussion

Table 4 reports performance on the protected test partition for relevance classification, strict span-based NER, multi-label tactic prediction, and technique- and sub-technique-level ATT&CK mapping. No baseline performs best across all tasks. Single-task supervised encoders are strongest for relevance and tactic prediction, while GPT-4o-mini achieves the best strict NER and fine-grained TTP-mapping performance.

Relevance classification. The single-task encoders provide the strongest relevance-classification results. SecBERT achieves the highest relevant-class F1 of 0.7585, whereas RoBERTa obtains the highest accuracy of 0.8116. SecBERT therefore provides a modest F1 advantage, while RoBERTa retains a small accuracy advantage, indicating that the effect of domain-specific pretraining is metric-dependent. Both MT-RoBERTa and GPT-4o-mini perform below the single-task encoders on this task.

Named Entity Recognition. Strict NER is more challenging than relevance classification. GPT-4o-mini achieves the highest exact-match F1 of 0.2233, followed by RoBERTa at 0.2074, MT-RoBERTa at 0.1753, and SecBERT at 0.1414. Thus, domain-specific pretraining does not consistently improve boundary-sensitive extraction. Because correct predictions require matching both entity type and exact character offsets, partial and boundary-misaligned spans are counted as errors.

Tactic prediction. SecBERT performs best on tactic prediction, reaching 0.5575 micro-F1 and 0.5113 macro-F1, followed by RoBERTa with 0.5261 and 0.4717, respectively. The consistent gain across both metrics suggests that cybersecurity-domain pretraining is useful for predicting the higher-level adversarial objectives represented by the 14 Enterprise ATT&CK tactics.

Technique and sub-technique mapping. Fine-grained ATT&CK mapping remains the most difficult task. GPT-4o-mini achieves the highest full-space performance, with a micro-F1 of 0.1344 and a macro-F1 of 0.0469. RoBERTa and SecBERT obtain similar micro-F1 values of 0.0570 and 0.0583, respectively, while MT-RoBERTa reaches 0.0090. The large micro-macro gap indicates uneven label-wise performance, with many identifiers receiving little or no correct recovery. The

training-frequency analysis in Section 5.5 further shows that training support alone does not explain the observed variation.

The supervised TTP classifiers are restricted to identifiers observed during training and therefore cannot recover gold assignments associated with test-occurring train-unseen identifiers. GPT-4o-mini is the only evaluated baseline to recover any such assignments, correctly predicting 4 of the 81 gold assignments. A detailed train-seen and train-unseen analysis is provided in Section 5.5.

Overall finding. The results show a task-dependent performance pattern rather than a single dominant model. Among the supervised encoders, SecBERT achieves the strongest relevant-class F1 and tactic-prediction results, but it does not provide consistent gains for strict NER or fine-grained TTP mapping. GPT-4o-mini performs best on strict NER and full-space TTP mapping, although its absolute scores remain low. These findings indicate that exact entity grounding and long-tailed ATT&CK mapping remain the principal challenges in SMAH-CTI.

5.5 Challenge Analysis

We analyze complementary challenges in SMAH-CTI, including the long-tailed ATT&CK label distribution, train-unseen identifiers, conditional relevance, multi-label prediction, and strict character-span grounding. Dataset-level statistics are computed from the frozen SMAH-CTI ground truth, while model-specific diagnostics use frozen GPT-4o-mini predictions on the complete protected test partition. No retraining, threshold tuning, or additional inference was performed.

Long-tail ATT&CK Distribution. Fine-grained ATT&CK mapping operates over a highly imbalanced label space. Across the dataset, 4,837 technique and sub-technique assignments span 383 ATT&CK identifiers. Median support is four assignments, and 229 identifiers (59.8%) occur five times or fewer. The 15 most frequent identifiers account for 36.3% of assignments, leaving 63.7% distributed across the remaining 368. Thus, much of the evaluation space receives limited supervision. This long-tail is reflected in the micro- and macro-F1 gap: micro-F1 aggregates decisions across labels while macro-F1 weights each label equally, including sparsely represented identifiers.

Train-Seen vs. Train-Unseen Identifiers. The supervised TTP classifiers use a closed vocabulary of 335 ATT&CK identifiers observed in training,

Table 4: Baseline performance on the protected SMAH-CTI test partition.

Model	Rel. F1	Rel. Acc.	NER F1	Tactic- μ F1	Tactic-MF1	TTP- μ F1	TTP-MF1
RoBERTa	0.7484	0.8116	0.2074	0.5261	0.4717	0.0570	0.0022
SecBERT	0.7585	0.8083	0.1414	0.5575	0.5113	0.0583	0.0007
MT-RoBERTa	0.7007	0.7976	0.1753	0.4221	0.2717	0.0090	0.0055
GPT-4o-mini	0.7013	0.7495	0.2233	0.4243	0.3686	0.1344	0.0469

Note: Rel.: relevance classification; NER: strict entity-type and character-offset matching; Tactic: multi-label tactic prediction; TTP: technique/sub-technique mapping. μ F1 and MF1 denote micro- and macro-F1. Bold indicates the best result in each column.

leaving 48 of the 383 evaluation identifiers train-unseen. Of these, 41 occur in the protected test partition, contributing 81 gold sentence-label assignments; the remaining seven have no test support. Because supervised classifiers cannot predict outside their training vocabulary, they cannot recover these 81 assignments.

GPT-4o-mini is the only baseline with an unrestricted output. It correctly predicts 4 of the 81 train-unseen assignments, yielding 4.94% recall. Over all 48 train-unseen identifiers, it produces 4 true positives, 6 false positives, and 77 false negatives, for a micro-F1 of 0.0879, compared with 0.1380 over the 335 train-seen identifiers.

Conditional Relevance Structure. SMAH-CTI follows a conditional structure in which entity, tactic, and technique annotations are associated with sentences containing adversarial behavior. On the 2,431-sentence protected test partition, GPT-4o-mini produces 715 true-positive, 462 false-positive, 147 false-negative, and 1,107 true-negative relevance decisions. The 147 false negatives contain 505 gold entities, 152 tactic assignments, and 146 technique or sub-technique assignments, quantifying the gold information unavailable to downstream components in a strictly relevance-gated pipeline. This is not a causal pipeline experiment, as downstream baselines are evaluated independently rather than under oracle and predicted relevance gates.

Multi-label TTP Complexity. TTP mapping is multi-label because a sentence may express multiple adversarial behaviors. Table 5 compares GPT-4o-mini on 768 single-label and 73 multi-label TTP-eligible test sentences, containing 768 and 154 gold assignments, respectively. Micro-F1 is 0.1282 for single-label and 0.1617 for multi-label sentences, providing no evidence that multi-label sentences are harder by this metric. However, exact label-set recovery remains difficult: only 36 single-label sentences (4.69%) achieve an exact match, compared with 1 multi-label sentence (1.37%). Thus, jointly recovering all supported behaviors remains uncommon.

Table 5: TTP difficulty (with GPT-4o-mini). MLC: multi-label complexity; LTF: long-tail frequency.

Type	Subset	Unit	Gold	μ -F1
MLC	Single	768 sent.	768	.1282
	Multi	73 sent.	154	.1617
LTF	Train-unseen (0)	48 labels	81	.0879
	Rare (1–5)	221 labels	206	.1523
	Medium (6–20)	73 labels	197	.1402
	Frequent (> 20)	41 labels	438	.1298

Training-Frequency Bands. Training-frequency analysis shows that label support alone does not determine performance. The unseen, rare, medium, and frequent bands contain 48, 221, 73, and 41 identifiers, accounting for 81, 206, 197, and 438 test gold assignments, respectively. Their micro-F1 values are 0.0879, 0.1523, 0.1402, and 0.1298. Performance remains low and does not improve monotonically with training support, suggesting that frequency is only one source of difficulty. The remaining variation is consistent with fine-grained label distinctions, hierarchical technique-sub-technique structure, and incomplete recovery of supported label sets.

Exact-Span Entity Grounding. NER uses strict character-span and entity-type matching: a prediction is correct only when its type and start and end offsets match a gold entity. GPT-4o-mini achieves an overall F1 of 0.2233, with 844 true positives, 3,174 false positives, and 2,699 false negatives over 3,543 gold entities. Performance varies across entity types: ThreatActor achieves 0.3829 F1, MalwareTool 0.3369, Infrastructure_Indicator 0.2098, and Action 0.1388. Action is both the most frequent and lowest-performing type, with 227 exact matches, 1,577 false positives, and 1,239 false negatives. These results highlight the difficulty of grounding behavioral descriptions to exact character spans rather than merely recognizing their presence. Overall, SMAH-CTI combines several challenges: a large long-tailed output space, train-unseen identifiers, conditional relevance, incomplete multi-label set recovery, and strict span-level grounding. Addressing them re-

quires models to move beyond topical recognition toward evidence-grounded, exact-span, and fine-grained multi-label extraction.

5.6 Error Analysis

We analyze errors in the frozen GPT-4o-mini predictions on the complete protected test partition. The analysis reports prediction–gold discrepancies based on exact labels and character offsets, without attributing unverified linguistic or causal explanations to individual errors.

Relevance Classification. GPT-4o-mini produces 715 true positives, 462 false positives, 147 false negatives, and 1,107 true negatives, yielding 0.7013 F1 and 0.7495 accuracy. The 147 false negatives contain 505 gold entities, 152 tactic assignments, and 146 technique or sub-technique assignments that would be missed by a strictly relevance-gated system. Conversely, 462 false positives assign relevance to gold-irrelevant sentences. These errors may reflect implicit behavior, contextual dependence, lexical overlap, or annotation ambiguity, but the counts alone cannot distinguish among these causes.

Named-Entity Recognition. Strict NER evaluation yields 844 exact matches, 3,174 false-positive spans, and 2,699 false-negative spans. Under deterministic one-to-one pairing, exact type-and-boundary matches are paired first, followed by identical-boundary type errors, same-type overlaps, and different-type overlaps. This yields 792 boundary mismatches, 86 type errors, and 157 cases differing in type and boundaries. After pairing, 1,664 gold and 2,139 predicted entities remain unmatched. Performance also varies by entity type: ThreatActor achieves 0.3829 F1, followed by MalwareTool (0.3369), InfrastructureIndicator (0.2098), and Action (0.1388). Action has the largest gold support and the most false-positive and false-negative spans. Thus, strict NER errors include boundary mismatches, type errors, and unmatched gold and predicted entities.

TTP Mapping. The TTP error inventory covers all 841 TTP-eligible test sentences. GPT-4o-mini returns no ATT&CK identifier for 616 sentences, making omission the dominant failure mode. Only 37 sentences exactly match the gold identifier set. Among the remainder, 143 predictions are disjoint from the gold set, 31 contain all gold identifiers plus additional ones, 10 contain both missing and additional identifiers, and 4 recover only a subset of the gold labels. Across the complete 383-

label space, GPT-4o-mini achieves a micro-F1 of 0.1344 and macro-F1 of 0.0469. The lower macro-F1 reflects uneven label-wise performance, with many identifiers receiving little or no recovery.

Unseen-Label Errors. Of 48 train-unseen IDs, 41 occur in the protected test partition, contributing 81 gold assignments; the remaining seven have no test support. Restricting gold and predicted identifiers to this 48-identifier subset, GPT-4o-mini produces 4 true positives, 6 false positives, and 77 false negatives, yielding 0.4000 precision, 0.0494 recall, and 0.0879 micro-F1. Thus, only 4 of 81 test-occurring train-unseen assignments are recovered, indicating that an open output vocabulary alone is insufficient for reliable recovery of ATT&CK identifiers absent from training.

Single-Label and Multi-Label Set Recovery.

The single-label subset contains 768 sentences and gold assignments, with 66 true positives, 196 false positives, and 702 false negatives, yielding 0.1282 micro-F1. The 73 multi-label sentences contain 154 gold assignments, with 19 true positives, 62 false positives, and 135 false negatives, yielding 0.1617 micro-F1. Despite higher micro-F1, exact-set recovery is lower for multi-label sentences: 36 single-label sentences are predicted exactly, compared with only one multi-label sentence. Thus, micro-F1 and exact-set accuracy capture different aspects of TTP performance.

Interpretation. The discrepancies show SMAH-CTI evaluates more than topical or lexical recognition. Relevance requires distinguishing behavioral evidence from context; NER requires exact entity types and boundaries; and TTP mapping requires recovering fine-grained identifiers from a large, hierarchical, and imbalanced labels. Thus, models may recognize cybersecurity content while failing the benchmark’s evidence-grounded criteria.

6 Conclusion and Future Work

We introduced SMAH-CTI, a unified sentence-level benchmark for grounded CTI extraction from real-world threat reports. It provides human-annotated supervision for relevance classification, CTI-specific entity extraction, and fine-grained multi-label ATT&CK mapping, supporting reproducible evaluation of individual tasks and end-to-end CTI extraction pipelines. Future work includes additional sources, multilingual reports, document-level TTP reasoning, and richer relational and temporal annotations.

7 Limitations

SMAH-CTI has several limitations. First, although the dataset is constructed from multiple real-world CTI providers, it remains limited to a fixed set of English-language sources and reporting styles. Consequently, certain domains, threat landscapes, emerging attack patterns, and non-English intelligence sources may be underrepresented, potentially affecting generalization to unseen report formats or niche threat domains.

Second, the dataset preserves the naturally imbalanced and long-tailed distribution of ATT&CK tactics, techniques, and sub-techniques observed in operational CTI reporting. While this improves realism, it also makes learning rare behaviors more difficult and may bias models toward frequently observed identifiers.

Third, despite detailed annotation guidelines, calibration, adjudication, and inter-annotator agreement analysis, some degree of subjectivity remains in relevance classification, entity boundary identification, and fine-grained ATT&CK mapping, particularly for implicit adversarial behavior or ambiguous sentence contexts.

Fourth, SMAH-CTI intentionally adopts sentence-level annotation to provide explicit grounding between each annotated TTP and the sentence containing its primary supporting evidence. To preserve local narrative continuity without sacrificing sentence-level supervision, each annotated sentence is accompanied by its immediately preceding sentence in the `context_before` field. This local context can help resolve nearby discourse dependencies while maintaining reproducible sentence-level annotation and evaluation. However, some adversarial behaviors and ATT&CK mappings may require reasoning across multiple sentences or paragraphs. Such document-level cases are beyond the scope of the current benchmark, whose primary objective is grounded sentence-level CTI extraction. Rigorously quantifying the proportion of ATT&CK mappings requiring document-level reasoning would require a separate document-level annotation protocol and independently constructed gold standard. We therefore leave this analysis and benchmark extension to future work.

Fifth, the dataset adopts MITRE ATT&CK as the primary ontology for behavior representation. Although widely used, ATT&CK does not capture all possible threat behaviors and evolves over

time, which may require future updates to maintain alignment with newer framework versions.

Finally, the unified multi-task design introduces additional evaluation complexity because the benchmark combines heterogeneous tasks, including relevance classification, sequence labeling, multi-label ATT&CK mapping, and grounded procedure extraction. In end-to-end extraction settings, errors may propagate across interconnected tasks, making it more difficult to attribute downstream failures to individual components and directly compare modeling approaches.

8 Ethics Statement

SMAH-CTI is derived from publicly available CTI reports and is intended to support defensive cybersecurity research. Because CTI resources are inherently dual-use, structured representations of adversarial behavior could potentially be misused for offensive purposes. However, the dataset does not contain exploit code, credentials, private victim information, or unpublished incident details. Instead, it normalizes publicly available information into sentence-level annotations for research and benchmarking.

The primary goal of SMAH-CTI is to support reproducible evaluation of grounded CTI extraction systems, including relevance classification, NER, ATT&CK mapping, and grounded procedure extraction. We encourage responsible use of the dataset for defensive research, threat analysis, and security automation, and discourage its use for harmful or unauthorized activities.

References

- Basel Abdeen, Ehab Al-Shaer, Anoop Singhal, Latifur Khan, and Kevin W. Hamlen. 2023. [SMET: semantic mapping of CVE to att&ck and its application to cybersecurity](#). In *Data and Applications Security and Privacy XXXVII - 37th Annual IFIP WG 11.3 Conference, DBSec 2023, Sophia-Antipolis, France, July 19-21, 2023, Proceedings*, volume 13942 of *Lecture Notes in Computer Science*, pages 243–260. Springer.
- Md Tanvirul Alam, Dipkamal Bhusal, Le Nguyen, and Nidhi Rastogi. 2024. [Ctibench: A benchmark for evaluating llms in cyber threat intelligence](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, pages 50805–50825.

- Md Tanvirul Alam, Dipkamal Bhusal, Youngja Park, and Nidhi Rastogi. 2023. [Looking beyond iocs: Automatically extracting attack patterns from external CTI](#). In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2023, Hong Kong, China, October 16-18, 2023*, pages 92–108. ACM.
- Asad Ali and Min-Chun Peng. 2024. [TTPMapper: Accurate mapping of ttps from unstructured cti reports](#). In *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, pages 558–563. IEEE.
- APTnotes. n.d. Aptnotes: A publicly available repository of advanced persistent threat reports. <https://github.com/aptnotes/data>. GitHub repository, accessed 2025.
- JunJun Chen, Chengliang Gao, Fei Tang, Jiaxu Xing, Qianlong Xiao, Dongyang Zheng, and Jing Qiu. 2023. [Automatically identifying sentences with attack behavior from cyber threat intelligence reports](#). In *8th International Conference on Data Science in Cyberspace, DSC 2023, Hefei, China, August 18-20, 2023*, pages 491–498. IEEE.
- CrowdStrike. 2024. [Crowdstrike global threat report](#).
- Sofia Della Penna, Roberto Natella, Vittorio Orbinato, Lorenzo Parracino, and Luciano Pianese. 2025. [CTI-HAL: A human-annotated dataset for cyber threat intelligence analysis](#). In *IEEE European Symposium on Security and Privacy, EuroS&P 2025 - Workshops, Venice, Italy, June 30 - July 4, 2025*, pages 69–78. IEEE.
- Gabriele Esposito. 2025. LLMs in the SIEM Loop: A Contract-Based Framework for Threat Detection with an Evaluation on Windows Telemetry and MITRE ATT&CK Mapping. Master’s thesis, Politecnico di Torino.
- Anooja Joy, Madhav Chandane, Yash Nagare, and Faruk Kazi. 2025. [Threat intelligence extraction framework \(TIEF\) for TTP extraction](#). *J. Cybersecur. Priv.*, 5(3):63.
- James Kotsias, Atif Ahmad, and Rens Scheepers. 2023. [Adopting and integrating cyber-threat intelligence in a commercial organisation](#). *Eur. J. Inf. Syst.*, 32(1):35–51.
- Philipp Kuehn. 2025. [The notion of relevance in cybersecurity: A categorization of security tools and deduction of relevance notions](#). In *Proactive Cyber Threat Intelligence: Automating the Intelligence Cycle based on Open Sources*, pages 61–76. Springer Vieweg.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Valentine Legoy, Marco Caselli, Christin Seifert, and Andreas Peter. 2020. Automated retrieval of att&ck tactics and techniques for cyber threat reports. *arXiv preprint arXiv:2004.14322*.
- Mandiant. 2024. [Mandiant threat intelligence reports](#). Part of Google Cloud.
- Renato Marinho and Raimir Holanda. 2023. [Automated emerging cyber threat identification and profiling based on natural language processing](#). *IEEE Access*, 11:58915–58936.
- MITRE Corporation. 2025. MITRE ATT&CK version 18.1. <https://attack.mitre.org/resources/versions/>. Enterprise ATT&CK, accessed January 2026.
- Aziz Mohaisen, Omar Alrawi, and Manar Mohaisen. 2015. [AMAL: High-fidelity, behavior-based automated malware analysis and classification](#). *Computers & Security*, 52:251–266.
- OASIS. 2021a. Structured threat information expression (stix) version 2.1. <https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html>. OASIS Standard.
- OASIS. 2021b. Trusted automated exchange of indicator information (taxii) version 2.1. <https://docs.oasis-open.org/cti/taxii/v2.1/os/taxii-v2.1-os.html>. OASIS Standard.
- Vittorio Orbinato, Mariarosaria Barbaraci, Roberto Natella, and Domenico Cotroneo. 2022. [Automatic mapping of unstructured cyber threat intelligence: An experimental study: \(practical experience report\)](#). In *IEEE 33rd International Symposium on Software Reliability Engineering, ISSRE 2022, Charlotte, NC, USA, October 31 - November 3, 2022*, pages 181–192. IEEE.
- Wei Peng, Junmei Ding, Wei Wang, Lei Cui, Wei Cai, Zhiyu Hao, and Xiaochun Yun. 2026. [Ctium: A new benchmark dataset for cyber threat intelligence summarization](#). *Comput. Secur.*, 168:104928.
- Md. Rayhanur Rahman, Brandon Wroblewski, Quinn Matthews, Brantley Morgan, Timothy Menzies, and Laurie A. Williams. 2025. [Mining temporal attack patterns from cyberthreat intelligence reports](#). *Knowl. Inf. Syst.*, 67(10):8941–8981.
- Nanda Rani, Bikash Saha, Vikas Maurya, and Sandeep Kumar Shukla. 2023. [TTPHunter: Automated extraction of actionable intelligence as ttps from narrative threat reports](#). In *Proceedings of the 2023 Australasian Computer Science Week, ACSW 2023, Melbourne, VIC, Australia, 30 January 2023 - 3 February 2023*, pages 126–134. ACM.
- Nanda Rani, Bikash Saha, Vikas Maurya, and Sandeep Kumar Shukla. 2024. [Ttpxhunter: Actionable threat intelligence extraction as ttps from finished cyber threat reports](#). *Digit. Threat. Res. Pract.*, 5(4):37:1–37:19.

Pedro Santos, Rafael Abreu, Manuel JCS Reis, Carlos Seródio, and Frederico Branco. 2025. [A systematic review of cyber threat intelligence: The effectiveness of technologies, strategies, and collaborations in combating modern threats](#). *Sensors*, 25(14):4272.

SentinelOne. 2024. [Sentinelone threat research and intelligence](#).

Giuseppe Siracusano, Davide Sanvito, Roberto Gonzalez, Manikantan Srinivasan, Sivakaman Kamatchi, Wataru Takahashi, Masaru Kawakita, Takahiro Kakumaru, and Roberto Bifulco. 2023. [Time for aCTIon: Automated analysis of cyber threat intelligence in the wild](#). *CoRR*, abs/2307.10214.

Mahzabin Tamanna, Shaswata Mitra, Md Erfan, Ahmed Ryan, Sudip Mittal, Laurie Williams, and Md Rayhanur Rahman. 2026. What are adversaries doing? automating tactics, techniques, and procedures extraction: A systematic review. *arXiv preprint arXiv:2604.02377*.

Xiuzhang Yang, Ruijie Zhong, Yuling Chen, Guojun Peng, Di Yao, Chaofan Chen, Chenyang Wang, Dongni Zhang, Yilin Zhou, and Zixuan Yang. 2026. [Cti-thinker: an llm-driven system for CTI knowledge graph construction and attack reasoning](#). *Cybersecur.*, 9(1):106.

Yizhe You, Jun Jiang, Zhengwei Jiang, Peian Yang, Baoxu Liu, Huamin Feng, Xuren Wang, and Ning Li. 2022. [TIM: threat context-enhanced TTP intelligence mining on unstructured threat data](#). *Cybersecur.*, 5(1):3.

Yongheng Zhang, Tingwen Du, Yunshan Ma, Xiang Wang, Yi Xie, Guozheng Yang, Yuliang Lu, and Ee-Chien Chang. 2025. [Attackg+: Boosting attack graph construction with large language models](#). *Comput. Secur.*, 150:104220.

A Annotation Granularity

Intermediate representations, such as paragraph-level or sliding-window formulations, partially mitigate document-level limitations but often combine multiple attack stages or shifts between operational and descriptive content, introducing mixed supervision signals (Esposito, 2025). This makes it difficult to assign coherent relevance labels or consistent multi-label ATT&CK mappings without introducing heuristic assumptions.

Span-level annotation enables precise extraction of entities and local patterns but may fail to capture complete semantic units of adversarial behavior. In CTI narratives, meaningful attacker actions frequently arise from interactions among entities and contextual cues distributed across a sentence rather than isolated spans.

Sentence-level annotation therefore offers an effective trade-off between semantic precision, explicit grounding, and modeling flexibility for end-to-end CTI extraction.

B Dataset Construction Details

B.1 Additional Preprocessing Details

Detailed preprocessing rules include curated regular-expression filters for removing pagination artifacts, headers, footers, legal disclaimers, contact information, and figure or table captions. Semantic line reconstruction is applied to repair artificial PDF line breaks and fragmented sentences introduced during text extraction.

Whitespace normalization and layout-aware parsing are additionally applied to improve sentence reconstruction quality across heterogeneous report formats. IoC-heavy tables and isolated indicator blocks are filtered when they lack explicit operational or behavioral context.

B.2 Annotation and Agreement

Detailed annotation guidelines define relevance criteria, entity-boundary rules, ATT&CK mapping conventions, and grounded procedure-writing policies. Ambiguous cases are resolved through iterative calibration meetings and adjudication by a senior reviewer.

Detailed inter-annotator agreement tables, confusion matrices, disagreement analyses, and entity-level precision/recall statistics are provided in the supplementary material.

B.3 Additional Dataset Statistics

Additional statistics, including extended ATT&CK coverage analyses, tactic–technique marginal distributions, long-tail frequency plots, and supplementary co-occurrence analyses, are provided in the supplementary material together with release resources and benchmark files.

C Additional Experimental Details

C.1 Baseline Implementation Details

For NER, RoBERTa is trained as a BIO token-classification model over the four SMAH-CTI entity types: THREATACTOR, MALWARETOOL, INFRASTRUCTURE_INDICATOR, and ACTION. Since standard BIO tagging cannot represent overlapping spans, overlapping entities are deterministically flattened for this baseline.

For TTP mapping, RoBERTa is trained as a multi-label classifier over ATT&CK technique and sub-technique identifiers observed in the training split. A binary cross-entropy objective with positive class weighting is used to mitigate label imbalance, and the decision threshold is selected on the development set.

For GPT-4o-mini prompting, extracted entity spans are validated against the source sentence using exact character offsets. For TTP mapping, outputs are normalized to ATT&CK technique and sub-technique identifiers before evaluation.

C.2 Additional Error Examples

Representative qualitative GPT-4o-mini error examples, including relevance-gating failures, exact-span grounding mismatches, and ATT&CK technique confusion cases, are provided in the supplementary material.

D Dataset Construction Details

D.1 Additional Preprocessing Details

Noise filtering removes pagination artifacts, corporate headers and footers, legal disclaimers, contact details, and figure or table captions using curated regular-expression rules. IoC filtering removes standalone indicators lacking explicit behavioral context while preserving operationally meaningful entries associated with adversarial activity. Preprocessing also includes whitespace normalization and semantic line reconstruction to repair artificial PDF line breaks and fragmented sentences introduced during text extraction.

D.2 Sentence-Level Modeling

Document-level annotations aggregate heterogeneous content, while paragraph-level representations often combine multiple attack stages or semantically distinct behaviors. Sentence-level processing enables explicit grounding of relevance labels, entities, procedures, and ATT&CK mappings to localized semantic units while supporting conditional extraction and fine-grained error analysis.

D.3 Additional Annotation Details

All entities are represented using span-based annotations with explicit character offsets. The ACTION category captures semantically complete adversarial verb phrases rather than isolated verbs.

The multi-label TTP formulation captures the compositional nature of adversarial behavior,

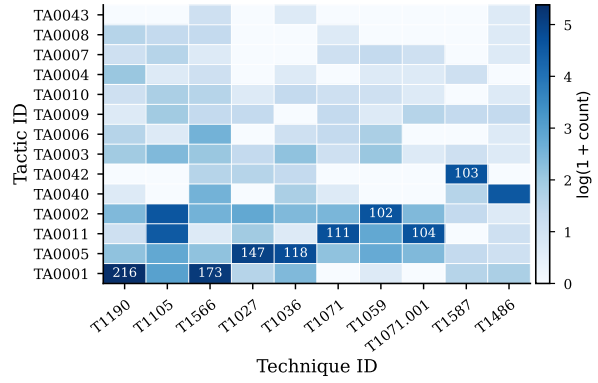


Figure 4: Sentence-level tactic-technique co-occurrence heatmap for the top-10 ATT&CK techniques in SMAH-CTI.

where a single sentence may describe overlapping attack objectives and procedures. Grounded procedure annotations preserve contextual and operational details that may not be fully represented by standardized ATT&CK identifiers alone.

D.4 Additional Dataset Statistics

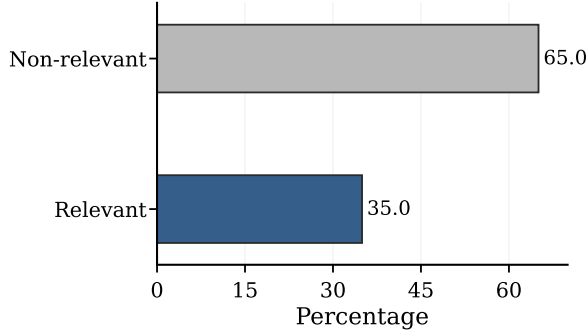
The dataset covers 383 unique ATT&CK identifiers, including 195 base technique families and 196 sub-techniques. Overall coverage includes 187/216 Enterprise ATT&CK base techniques and 196/475 sub-techniques in ATT&CK v18.1.

Figure 4 illustrates realistic multi-stage adversarial behavior, where techniques associated with initial compromise, obfuscation, execution, and command-and-control frequently co-occur across tactical objectives.

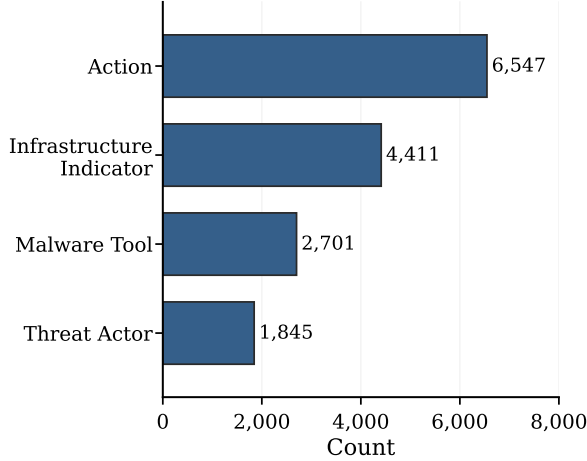
Figure 5a shows that 35.0% of sentences are labeled as behaviorally relevant, reflecting the naturally imbalanced composition of operational CTI reports. Figure 5b presents the distribution of entity types, where ACTION entities are the most frequent, followed by INFRASTRUCTURE_INDICATOR, MALWARE_TOOL, and THREAT_ACTOR.

D.5 Release and Reproducibility

The final ground truth dataset contains 12,304 validated sentence-level instances and is released in JSON and CSV formats with deterministic identifiers and preserved provenance metadata. Exact report identifiers assigned to each split, together with evaluation scripts and release resources, are distributed with the benchmark to support transparent and reproducible evaluation.



(a) Relevance label distribution.



(b) Entity type distribution.

Figure 5: Distribution of relevance labels and entity types in SMAH-CTI.

E Supported Tasks

SMAH-CTI is designed as a unified, sentence-level benchmark for structured CTI extraction. By integrating relevance classification, NER, and fine-grained MITRE ATT&CK mapping in a single framework, it supports both isolated task evaluation and end-to-end CTI extraction pipelines.

E.1 SMAH-CTI-Enabled Questions

SMAH-CTI enables investigation of research questions that are difficult to study with prior CTI datasets lacking sentence-level grounding or unified multi-task supervision. **❶ How does relevance classification affect downstream CTI extraction?** Since NER and TTP mapping are applied to behaviorally relevant sentences, relevance errors may propagate to downstream extraction stages. **❷ How can models represent multi-label adversarial behavior?** CTI sentences may express multiple tactics, techniques, or sub-techniques simultaneously, requiring fine-grained multi-label reasoning under long-tail label distributions. **❸ How can grounded extraction**

reduce hallucination risk? SMAH-CTI links structured outputs to explicit sentence-level evidence, enabling evaluation of whether extracted entities, procedures, and ATT&CK mappings are textually supported. **❹ Can cross-task learning improve CTI extraction?** The dataset supports analysis of dependencies between relevance classification, NER, and ATT&CK mapping in unified pipelines.

E.2 Supported Tasks

SMAH-CTI supports the following tasks: **❶ Relevance Classification:** Binary classification of sentences based on the presence of operational adversarial behavior, serving as a filtering stage for downstream extraction. **❷ NER:** Extraction of CTI-specific entities, including THREATACTOR, MALWARETOOL, INFRASTRUCTUREINDICATOR, and ACTION, enabling structured representation of adversarial behavior. **❸ TTP Mapping:** Multi-label mapping of relevant sentences to ATT&CK tactics, techniques, and sub-techniques, together with grounded procedure descriptions linked to source text. **❹ End-to-End and Conditional CTI Extraction:** Unified modeling of relevance classification, entity extraction, and ATT&CK mapping within integrated CTI extraction pipelines, where relevance classification acts as a gating mechanism for downstream structured extraction.

E.3 Potential Applications

Beyond benchmarking, SMAH-CTI supports several research and operational applications: **❶ Automated Structuring of CTI Reports:** Transforming unstructured threat reports into structured, machine-readable intelligence for scalable indexing, retrieval, and analysis. **❷ Behavior-Centric Threat Analysis:** Supporting ATT&CK-grounded analysis of adversarial behavior for detection engineering, threat hunting, and defense planning. **❸ Threat Knowledge Graph Construction:** Enabling structured modeling of relationships among adversaries, tools, infrastructure, and techniques for reasoning and campaign analysis. **❹ Security Operations Support:** Facilitating threat hunting, incident investigation, and alert enrichment through structured ATT&CK-aligned extraction grounded in source text. **❺ Standardized Benchmarking and Evaluation:** Providing a unified benchmark for reproducible comparison across rule-based, machine learning, neural, and

Table 6: Core fields in the released SMAH-CTI annotation schema.

Field	Description
uid	Deterministic sentence identifier derived from provenance metadata and a short text hash.
source	Source or provider of the CTI report.
doc_id	Document-level identifier for the source report.
page_number	Page number or normalized page index associated with the sentence.
sentence_id	Sentence index preserving document order and traceability.
context_before	Immediately preceding sentence retained for local narrative context.
sentence_text	Sentence annotated in the ground truth.
relevant	Binary label indicating whether the sentence contains operational adversarial behavior.
tactic	List of ATT&CK tactic objects; null for irrelevant sentences.
techniques	List of ATT&CK technique or sub-technique objects; null for irrelevant sentences.
procedure	Grounded free-text description of the behavior; null for irrelevant sentences.
entities	List of span-based entity annotations; null for irrelevant sentences.

LLM-based CTI extraction systems.

F Annotation Format

This appendix documents the SMAH-CTI annotation format and supporting methodological details. To support reproducibility and programmatic use, Section F.1 summarizes the core fields in the final ground truth, Section F.2 formalizes the annotation workflow, Section F.3 describes the annotation constraints and consistency rules, Section F.4 reports detailed inter-annotator agreement results, Section F.5 presents representative sentence-level annotation examples, and Section F.6 provides representative model error examples from the GPT-4o-mini test-set outputs.

F.1 Annotation Schema

Table 6 summarizes the core fields used in the released SMAH-CTI ground truth. Each instance corresponds to a single sentence and includes metadata, textual content, and conditionally defined task-specific annotations. Task-specific fields (*e.g.*, entities, tactics, techniques, and procedure) are populated only for sentences labeled as relevant, ensuring consistency and preventing noise from non-operational content.

Algorithm 1 Sentence-level annotation workflow for constructing SMAH-CTI.

Require: CTI reports $\mathcal{D}$
Ensure: Annotated dataset $\mathcal{S}$

- 1: Preprocess documents (cleaning, noise removal)
- 2: Segment documents into sentences $\{s_i\}$ ▷
- Sentence-level grounding
- 3: **for** each sentence s_i **do**
- 4: Assign relevance label $r_i \in \{\text{yes}, \text{no}\}$ ▷ Behavioral evidence
- 5: **if** $r_i = \text{yes}$ **then**
- 6: Annotate entity spans ▷ Exact character boundaries
- 7: Assign multi-label ATT&CK tactics $\mathcal{T}_i$
- 8: Assign multi-label ATT&CK techniques/sub-techniques $\mathcal{C}_i$ ▷ Evidence-grounded mapping
- 9: Write grounded procedure p_i ▷ Instance-specific behavior
- 10: **else**
- 11: Set $\mathcal{T}_i, \mathcal{C}_i, p_i$, and entities to null ▷ Enforce cross-layer consistency
- 12: **end if**
- 13: Store structured annotation for s_i
- 14: **end for**
- 15: **return** $\mathcal{S}$

F.2 Annotation Workflow and Construction

The annotation process follows a structured, multi-stage pipeline designed to ensure consistency, reproducibility, and explicit grounding in textual evidence. Algorithm 1 summarizes this sentence-level annotation workflow.

F.3 Annotation Constraints and Consistency

To ensure semantic coherence, annotation quality, and reproducibility, the SMAH-CTI annotation process follows a set of explicit constraints and consistency rules derived from the dataset construction and normalization pipeline.

Relevance-Annotation Dependency. Task-specific annotations are conditionally defined based on sentence-level relevance. Only sentences labeled as relevant contain structured annotations (entities, tactics, techniques, and procedure), while all task-specific fields are set to `null` for irrelevant sentences. This constraint ensures that structured outputs are grounded exclusively in operational adversarial behavior and prevents noise from non-operational content.

Behavioral Grounding Constraint. A sentence is annotated as relevant only if it explicitly describes observable adversarial actions that can be mapped to at least one MITRE ATT&CK tactic. Sentences containing only contextual information, standalone indicators of compromise, mitigation guidance, or background analysis are labeled as ir-

relevant. This enforces alignment between textual evidence and ATT&CK-based representation.

Span-Level Entity Consistency. Named entities are annotated as contiguous spans with explicit character offsets and predefined labels (THREATACTOR, MALWARETOOL, INFRASTRUCTURE_INDICATOR, and ACTION). Entity annotations must correspond to semantically meaningful units within the sentence and remain consistent with the described adversarial behavior. Overlapping or semantically redundant spans are avoided to maintain annotation clarity.

Multi-Label TTP Alignment. Each relevant sentence may be assigned multiple tactics and techniques (including sub-techniques), reflecting the compositional nature of behavior. All ATT&CK mappings must use official identifiers and names from the Enterprise ATT&CK framework to ensure standardization and interoperability.

Procedure Grounding Constraint. The `procedure` field provides a free-text, sentence-level description of the adversarial behavior. It must remain semantically consistent with the annotated entities and ATT&CK mappings, capturing the concrete realization of the behavior without introducing information not present in the original sentence.

Cross-Field Logical Consistency. To enforce internal coherence, the following constraints are applied during normalization: (i) irrelevant sentences must not contain entities, tactics, techniques, or procedures, (ii) relevant sentences must contain at least one valid tactic-level ATT&CK mapping when behavior can be aligned with ATT&CK, while technique-level labels may be `null` when the sentence lacks sufficient operational detail, and (iii) entity annotations must align with the behavior described in the `procedure`. Any violations are deterministically corrected during post-processing.

Provenance and Traceability Constraint. Each annotated instance preserves document-level provenance through `doc_id`, `page_number`, and `sentence_id`, and is uniquely identified using a deterministic `uid`. This ensures traceability to the original source and supports reproducibility across dataset versions.

Context Preservation Constraint. Each sentence is associated with its immediately preceding sentence (`context_before`) to maintain local narrative continuity. This provides minimal contextual grounding while preserving sentence-level in-

dependence for annotation and evaluation.

Normalization and Schema Consistency. All annotations are normalized into a canonical schema with standardized field formats, ATT&CK-aligned identifiers, and consistent ordering across instances. This ensures compatibility with downstream processing pipelines and enables reproducible evaluation.

F.4 Inter-Annotator Agreement Details

This subsection provides the detailed agreement results corresponding to Section 4.5. All agreement results are computed on the same 1,000-sentence controlled overlap subset. For sentence-level relevance, the annotators achieved an accuracy of 89.80% and Cohen’s Kappa of $\kappa = 0.751$, indicating substantial agreement according to Landis & Koch (Landis and Koch, 1977). The corresponding confusion matrix is shown in Table 7. For entity annotation, agreement is reported using presence-based precision, recall, and F1-score for each entity type. As shown in Table 9, precision was consistently high across entity types, whereas recall was lower for THREATACTOR, reflecting the frequent implicitness of actor mentions in CTI narratives. For ATT&CK mapping, agreement was computed using exact match and mean Jaccard similarity. Technique-level agreement reached 81.5% exact match with mean Jaccard similarity of 0.816, and tactic-level agreement reached 85.1% exact match with mean Jaccard similarity of 0.852. These results are summarized in Table 8. A total of 102 relevance disagreement cases were subsequently reviewed by a third adjudicator. The adjudicator examined the sentence text, surrounding context, and both annotators’ labels before assigning the final gold annotation according to the SMAH-CTI annotation guidelines. For sentences adjudicated as relevant, the associated structured annotations, including tactics, techniques, procedures, and entities, were validated and corrected where necessary. For sentences adjudicated as non-relevant, all structured fields were normalized to null values to preserve schema consistency. The reported IAA metrics correspond to the pre-adjudication overlap analysis, whereas the adjudicated labels were incorporated into the final released SMAH-CTI ground-truth dataset.

Table 7: Confusion matrix for sentence-level relevance agreement with annotators A \ B.

A \ B	Non-relevant	Relevant
Non-relevant	232	7
Relevant	95	666

Table 8: ATT&CK tactic and technique agreement.

Layer	Exact Match (%)	Mean Jaccard
Technique	81.5	0.816
Tactic	85.1	0.852

F.5 Illustrative Annotation Examples

The following examples illustrate the conditional structure of the SMAH-CTI annotation schema under both relevant and non-relevant settings. They show how task-specific annotations are instantiated for behaviorally relevant sentences and set to null for non-relevant sentences.

Example A1. Relevant sentence with structured multi-task annotation.

Table 9: Presence-based agreement for CTI-specific entity annotation.

Entity Type	Precision	Recall	F1-score
THREATACTOR	0.979	0.442	0.609
MALWARETOOL	0.967	0.723	0.827
INFRASTRUCTURE_INDICATOR	0.976	0.672	0.796
ACTION	0.988	0.632	0.771

```
{
  "uid": "mitre-96_mitre_report-pl-sj
  ↪ 40-22cb8b",
  "source": "mitre",
  "doc_id": "96_mitre_report",
  "page_number": 1,
  "sentence_id": 40,
  "context_before": ", ROADSWEET will
  ↪ begin resolving the necessary
  ↪ APIs using the Windows
  ↪ GetProcAddress API.",
  "sentence_text": "ROADSWEEP
  ↪ contains multiple embedded
  ↪ scripts which are used to
  ↪ either execute additional
  ↪ commands or to remove itself
  ↪ from the victim's device.",
  "relevant": "yes",
  "tactic": [
    {"id": "TA0002", "name":
    ↪ "Execution"},
    {"id": "TA0005", "name": "Defense
    ↪ Evasion"}
  ],
  "techniques": [
    {"id": "T1070.004", "name": "File
    ↪ Deletion"}
  ],
  "procedure": "ROADSWEEP executes
  ↪ commands and removes itself
  ↪ from the victim's device using
  ↪ embedded scripts.",
  "entities": [
    {"text": "ROADSWEEP", "start": 0,
    ↪ "end": 9, "label":
    ↪ "MalwareTool"},
    {"text": "execute additional
    ↪ commands", "start": 70,
    ↪ "end": 97, "label":
    ↪ "Action"},
    {"text": "remove itself from the
    ↪ victim's device", "start":
    ↪ 105, "end": 142, "label":
    ↪ "Action"}
  ]
}
```

Example A2. Irrelevant sentence with null task-specific fields.

```
{
  "uid": "mitre-89_mitre_report-p1-s_
  ↪ 144-69d73a",
  "source": "mitre",
  "doc_id": "89_mitre_report",
  "page_number": 1,
  "sentence_id": 144,
  "context_before": "Use the right
  ↪ tools - The technical burden of
  ↪ securing open source is
  ↪ yours.",
  "sentence_text": "Don't rely on the
  ↪ maintainers, there are tools
  ↪ that can help you protect your
  ↪ production workloads from the
  ↪ risks of using open source in
  ↪ runtime.",
  "relevant": "no",
  "tactic": null,
  "techniques": null,
  "procedure": null,
  "entities": null
}
```

Example A3. Relevant sentence with tactic-level annotation only. This example shows behavior mappable to a high-level ATT&CK tactic but not to a specific technique or sub-technique.

```
{
  "uid": "crowdstrike-46_crowdstrike_
  ↪ _report-p31-s276-973f28",
  "source": "crowdstrike",
  "doc_id": "46_crowdstrike_report",
  "page_number": 31,
  "sentence_id": 276,
  "context_before": "However, some
  ↪ threat actors are using
  ↪ AITM-based phishing, which
  ↪ proxies the phished user's
  ↪ authentication requests to the
  ↪ cloud authentication service
  ↪ being targeted.",
  "sentence_text": "This allows the
  ↪ threat actor to prompt the user
  ↪ for an MFA token as well,
  ↪ circumventing one of the
  ↪ primary cloud account security
  ↪ controls.",
  "relevant": "yes",
  "tactic": [
    {"id": "TA0001", "name": "Initial
    ↪ Access"}
  ],
  "techniques": null,
  "procedure": "The threat actor
  ↪ prompts the user for an MFA
  ↪ token to bypass cloud account
  ↪ security controls.",
  "entities": [
    {"text": "threat actor", "start":
    ↪ 16, "end": 28, "label":
    ↪ "ThreatActor"},
    {"text": "prompt the user for an
    ↪ MFA token", "start": 32,
    ↪ "end": 64, "label": "Action"}
  ]
}
```

Example A4. Multi-step sentence with multi-

label adversarial behavior.

```
{
  "uid": "aptnotes-02_aptnotes_repor_
  ↪ t-p2-s38-a6f674",
  "source": "aptnotes",
  "doc_id": "02_aptnotes_report",
  "page_number": 2,
  "sentence_id": 38,
  "context_before": "Producing
  ↪ Corporate Promotional
  ↪ Materials",
  "sentence_text": "By attaching
  ↪ files or including download
  ↪ links in the emails, the threat
  ↪ actor prompted users to
  ↪ download the compressed file
  ↪ and execute the LNK shortcut
  ↪ file inside.\n2/20",
  "relevant": "yes",
  "tactic": [
    {"id": "TA0001", "name": "Initial
    ↪ Access"},
    {"id": "TA0002", "name":
    ↪ "Execution"}
  ],
  "techniques": [
    {"id": "T1566", "name":
    ↪ "Phishing"},
    {"id": "T1204", "name": "User
    ↪ Execution"}
  ],
  "procedure": "Attach files or
  ↪ include links in emails to
  ↪ prompt users to download and
  ↪ execute a malicious LNK file.",
  "entities": [
    {"text": "attaching files or
    ↪ including download links in
    ↪ the emails", "start": 3,
    ↪ "end": 60, "label":
    ↪ "Action"},
    {"text": "the threat actor",
    ↪ "start": 62, "end": 78,
    ↪ "label": "ThreatActor"},
    {"text": "prompted users to
    ↪ download the compressed file
    ↪ and execute the LNK shortcut
    ↪ file inside", "start": 79,
    ↪ "end": 166, "label":
    ↪ "Action"},
    {"text": "emails", "start": 54,
    ↪ "end": 60, "label":
    ↪ "Infrastructure_Indicator"},
    {"text": "the compressed file",
    ↪ "start": 106, "end": 125,
    ↪ "label":
    ↪ "Infrastructure_Indicator"},
    {"text": "LNK shortcut file",
    ↪ "start": 142, "end": 159,
    ↪ "label": "MalwareTool"}
  ]
}
```

F.6 Representative Model Errors

The following examples present representative GPT-4o-mini prediction errors selected directly from the SMAH-CTI test-set prediction outputs. They illustrate the recurring error patterns dis-

cussed in Section 5.6. These examples are not intended as an exhaustive error taxonomy, but as concrete cases showing how relevance classification, exact-span NER, and ATT&CK technique mapping can fail under realistic CTI language.

Example A5. Relevance false positive.

uid: aptnotes-04_aptnotes_report-p1-s1-952e6c.

Error type: False positive.

Sentence: *“From Albania to the Middle East: The Scarred Manticore is Listening research.checkpoint.com/2023/from-albania-to-the-middle-east-the-scarred-manticore-is-listening October 31, 2023 Key Findings Check Point Research (CPR) is monitoring an ongoing Iranian espionage campaign by Scarred Manticore, an actor affiliated with the Ministry of Intelligence and Security (MOIS).”*

Gold label: non-relevant.

Model prediction: relevant.

Analysis: The sentence contains CTI-related context, including campaign and actor information, but it does not describe a concrete adversarial action. The model overgeneralizes from topical CTI language and incorrectly treats contextual threat reporting as behaviorally grounded relevance.

Example A6. NER missed entity extraction.

uid: aptnotes-04_aptnotes_report-p4-s32-4289b5

Error type: Missed entity extraction

Sentence: *“Configuration. The malware starts by performing a one-byte XOR decryption of a structure containing the malware configuration, which is represented with the following structure:”*

Gold annotation: ACTION: starts by performing a one-byte XOR decryption

Model prediction: No entity predicted

Analysis: The sentence describes a concrete malware behavior involving XOR decryption of configuration data, but the model fails to extract the gold ACTION span. This illustrates a recall-oriented NER error in which behaviorally meaningful CTI content is present but not grounded as an entity span.

Example A7. TTP related-technique confusion.

uid: aptnotes-04_aptnotes_report-p2-s20-clab67

Error type: Related-technique confusion

Sentence: *“It is a lightweight but rather sophisticated passive backdoor installed on Windows servers that enables attackers to execute commands remotely through HTTP requests.”*

Gold technique: T1071.001

Model prediction: T1059

Analysis: The prediction is behaviorally plausible because the sentence mentions remote command execution, but the gold annotation emphasizes communication through HTTP requests. This illustrates how models may select a plausible but unsupported ATT&CK technique when a sentence contains both command-execution language and protocol-specific communication evidence.