Making the Most of Fragmented Supervision for Sentence-to-TTP Mapping

1st Anonymous

***Abstract*—Mapping cyber threat intelligence (CTI) reports to MITRE ATT&CK techniques remains challenging. Although recent work has proposed increasingly sophisticated models for sentence-to-TTP classification, progress remains limited not only by the scarcity of high-quality labeled data, but also by the fragmentation of existing supervision across heterogeneous annotation granularities and label spaces, leaving much of it unusable for sentence-level learning.**

In this paper, we systematically analyze datasets used in CTI-to-ATT&CK research and show how their supervision properties limit direct reuse in benchmark settings. Guided by this analysis, we propose two complementary strategies to recover and transfer supervision. First, we introduce a label *recovery* mechanism that combines hierarchy-aware mapping with embedding-based semantic matching to preserve supervision otherwise discarded by benchmark label restrictions. Second, we propose *TTP-SiamAlign*: a Siamese bi-encoder retrieval approach that *transfers* existing document-level CTI annotations into sentence-level training pairs by identifying technique-aligned sentences within labeled reports.

Across two benchmark datasets, the combined augmentation strategy improves sentence-to-TTP classification performance by 3.9–12.1%, demonstrating that better utilization of existing supervision sources can yield substantial gains without requiring new manual annotations [1].

***Index Terms*—link prediction, CWE, TTPs.**

I. Introduction

Cyber Threat Intelligence (CTI) reports contain valuable evidence about adversary behavior, but most of this information is embedded in unstructured natural language. To support downstream security operations such as threat hunting, detection engineering, and intelligence sharing, analysts often map report content to the MITRE ATT&CK framework [1], which provides a standardized vocabulary of tactics, techniques, and procedures (TTPs). Manual CTI-to-ATT&CK annotation, however, is labor-intensive, slow, and difficult to scale, motivating growing work on automated TTP extraction from CTI text [2]–[4].

Recent work has increasingly converged on *sentence-to-TTP mapping*, in which individual CTI sentences or short evidence snippets are assigned one or more ATT&CK techniques [5]–[8]. This formulation preserves the evidence-to-label link needed for analyst verification and downstream use, making it a practical focal point for benchmarking and model development. Yet progress in this setting is constrained less by model design alone than by supervision quality and availability: sentence-level TTP annotations are expensive to produce, sparse, and label distributions are strongly long-tailed. Moreover, existing CTI resources differ substantially in annotation granularity, coverage, and label space, fragmenting usable supervision and complicating reproducible evaluation across studies [2], [9], [10].

We argue that a central bottleneck in sentence-to-TTP extraction is therefore *supervision utilization*: how to recover, align, and transfer heterogeneous supervision sources into usable sentence-level training signal. Beyond benchmark sentence labels, potentially useful supervision already exists in ATT&CK text and in document-level CTI annotations, but these sources are often used in isolation or discarded due to label-space mismatch. In this paper, we present a data-centric framework that addresses this bottleneck through two complementary mechanisms: (i) label-space recovery, which uses ATT&CK hierarchy structure and semantic similarity to preserve supervision otherwise excluded by benchmark label restrictions, and (ii) retrieval-based grounding, which uses a Siamese bi-encoder to mine high-confidence sentence–TTP pairs from document-labeled CTI reports by aligning CTI sentences with ATT&CK technique descriptions in a shared embedding space.

We make the following contributions:

- **Diagnosis:** We characterize supervision sources used for CTI-to-ATT&CK mapping and quantify how annotation granularity and benchmark label restrictions discard otherwise usable supervision.
- **Recovery:** We propose label-space recovery mechanisms based on ATT&CK hierarchy structure and semantic similarity to reuse supervision that would otherwise be excluded under benchmark constraints.
- **Transfer:** We introduce *TTP-SiamAlign*, a retrieval-based grounding procedure that converts document-level CTI labels into pseudo-labeled sentence–TTP pairs by linking report-level labels to specific supporting sentences.

We first analyze supervision fragmentation across the CTI-to-ATT&CK dataset ecosystem (Section IV). We then present our supervision-utilization framework (Sec-

[1]Code at http:\\github.com

tion V) and evaluate its impact on TRAM and AnnoCTR through controlled ablations and benchmark comparisons (Sections VI).

II. Background and Problem Formulation

The MITRE ATT&CK framework provides a standardized knowledge base of adversarial behavior, organizing tactics, techniques, and procedures (TTPs) into a shared vocabulary widely used in threat hunting, detection engineering, intelligence analysis, and reporting. By describing attacker behavior in a consistent and operationally meaningful way, ATT&CK helps analysts compare campaigns, communicate findings, and connect observed activity to defensive actions.

A. Supervision Sources for CTI-to-ATT&CK Learning

CTI-to-TTP learning naturally exposes multiple forms of supervision with different strengths and limitations:

- **Sentence-level sentence–TTP annotations** provide direct, evidence-grounded supervision, but are expensive to produce and typically limited in scale.
- **Document-level TTP labels** indicate which techniques appear in a CTI report, but do not specify which sentences express each technique.
- **ATT&CK textual semantics** (e.g., technique descriptions and related ATT&CK text) provide natural-language definitions of labels that can support semantic matching and retrieval.

These supervision sources are heterogeneous but complementary: sentence-level labels provide precision, document-level labels provide broader coverage, and ATT&CK text provides semantic structure. However, existing approaches typically exploit only a subset of these signals, which limits the effective supervision available to sentence-level models.

B. Problem Setting

We study sentence-to-TTP extraction in a data-scarce setting, where only a limited set of expert-annotated sentence–TTP pairs is available, while additional supervision exists in weaker or differently structured forms, including document-level CTI labels and ATT&CK textual knowledge. We identify *label-space mismatch* and *ungrounded document labels* as the main supervision bottlenecks in this setting, and address them through a data-centric augmentation pipeline that recovers and transfers auxiliary supervision into usable sentence-level training signal.

More concretely, label-space mismatch arises when auxiliary supervision is expressed over ATT&CK techniques that fall outside benchmark-restricted label sets, causing potentially useful supervision to be discarded. Ungrounded document labels arise because document-level CTI annotations indicate which techniques occur in a report, but not which sentence expresses each technique, making direct reuse for sentence-level learning unreliable. These challenges motivate methods for recovering otherwise excluded supervision and grounding weak document-level labels to evidence sentences.

III. Related Work

CTI-to-ATT&CK research has produced supervision in several different forms, including token-level, sentence-level, and document-level label annotations, to TTPs [2], [4], [9]. Among these, sentence-to-TTP supervision has emerged as the most practical target for benchmarking and downstream analyst use, since it preserves the evidence-to-label link needed for verification, interpretation, and operational deployment. However, the broader supervision ecosystem remains highly heterogeneous: resources differ in annotation granularity, label coverage, and benchmark compatibility, which makes much of the available signal difficult to use directly for sentence-level learning.

A. Sentence-Level Methods and Data Augmentation

To address the scarcity of expert-annotated sentence–TTP pairs, prior sentence-level work has explored several augmentation and auxiliary-supervision strategies. A common line of work treats the task as multi-label text classification and improves performance through synthetic expansion of the training data. Kim *et al.* [11], for example, apply lexical augmentation techniques such as EDA and back-translation to increase variation in scarce classes. TTPXHunter [12] performs context-preserving lexical substitution by masking target tokens and generating candidate replacements with a masked language model, retaining only semantically similar variants. HM-CAT [13] instead uses LLM-based generation to synthesize additional CTI text for underrepresented categories.

A complementary direction augments sentence-level learning with ATT&CK-derived auxiliary text. You *et al.* [14], for example, leverage ATT&CK procedure text in a two-stage training pipeline, showing that technique-associated natural language can improve downstream classification. These methods demonstrate that augmentation and auxiliary supervision can mitigate data scarcity, but they typically operate within a single augmentation paradigm or supervision source. In contrast, our work focuses on systematically recovering and integrating supervision across multiple heterogeneous resources rather than introducing a single new augmentation recipe.

B. Recover and transfer Knowledge

Several works recognize that useful information for CTI-to-ATT&CK mapping exists beyond expert-labeled sentence pairs. One line of work incorporates ATT&CK

structure or textual knowledge directly into the learning process, for example through hierarchy-aware modeling [13] or by leveraging ATT&CK descriptions and procedures as auxiliary semantic context [8], [14]. These approaches improve label modeling and semantic alignment, but typically assume a fixed benchmark label space and do not explicitly recover supervision that would otherwise be discarded due to granularity or label-space mismatch. Another line of work considers weaker supervision sources, such as document-level CTI corpora. While such resources are easier to obtain than sentence-level annotations, they only provide report-level technique labels and do not specify which sentence expresses which technique, making direct supervision transfer difficult [15]. As a result, prior work has only partially exploited these resources for sentence-level learning.

Our work is closely related to this direction but differs in two key aspects: first, we treat label-space mismatch itself as a recoverable supervision problem; second, we use retrieval-based grounding to convert document-level CTI labels into pseudo-labeled sentence–TTP pairs.

IV. Diagnosis: Dataset Ecosystem and Characterization

To support a data-centric analysis of sentence-to-TTP learning, we first characterize the dataset ecosystem used in prior CTI-to-ATT&CK research. Our goal is not only to list datasets, but to understand *which supervision sources are actually used*, at what granularity, and with what implications for sentence-level learning.

A. Collection Scope and Inclusion Criteria

We compiled papers that either (i) propose a method for mapping CTI text to MITRE ATT&CK techniques or (ii) introduce a dataset used for CTI-to-ATT&CK mapping. We identified candidates through Google Scholar search queries and screened them by title and abstract, followed by backward/forward snowballing to recover relevant papers that were not captured in the initial search. To reduce the risk of missing post-publication releases, we also checked whether code or datasets were later published by searching for each paper title together with "GitHub".

A key design choice in our catalog is to represent datasets at the level of *atomic source blocks*. When a paper reports using an aggregated dataset assembled from multiple underlying sources (e.g., a composite corpus built from several report repositories), we record the constituent sources rather than only the aggregate name. This avoids double-counting and enables a more accurate view of reuse across the literature. For example, if a method reports using an aggregated corpus built from multiple CTI repositories, we characterize dataset usage in terms of those underlying repositories.

We also track whether each resource is reproducible in practice. A resource is marked *reproducible* if it can be directly downloaded or reconstructed (even partially) from the paper and accompanying artifacts; otherwise, it is marked *not reproducible* (e.g., missing retrieval details, dead links, or unspecified scraping/filtering procedures). Finally, we check the overlapping across datasources as further explained in Section XI-A, and present the full overview in Table IX in the Appendix.

B. Resource Taxonomy and Annotation Granularity

To understand which resources can support sentence-to-TTP learning, we characterize each dataset/source by annotation granularity and supervision properties, as shown in Table I. We distinguish among:

- **Word-/phrase-level resources**, where labels are attached to spans or extracted actions;
- **Sentence-level resources**, where sentences/snippets are mapped to one or more techniques;
- **Document-level resources**, where reports are annotated with a set of techniques without sentence grounding.

We further record properties that affect practical usability for sentence-to-TTP learning, including: the number of reports/documents (when applicable), number of sentences, ATT&CK label coverage (techniques and subtechniques), whether negative evidence is present (i.e., sentences with no TTP label), and whether the resource has been used in prior sentence-to-TTP literature. This characterization makes explicit that resources differ not only in size, but also in the *type of supervision they provide* and the assumptions required to use them.

TABLE I
Characterization of datasets for sentence to TTP classification

Dataset	Gran.	#CTRs	#Sent.	TA	T/ST	NE	Ratio	Used
AnnoCTR [16]	Word	120	4652	12	130	●	58.19%	●
TRAM [17]	Sent.	150	19,178	❍	50	●	21.22%	●
ATT&CK proc - malware	Sent.	❍	9836	0	418	❍	100%	◗
ATT&CK proc - group	Sent.	❍	4362	0	488	❍	100%	◗
ATT&CK proc - campaign	Sent.	❍	1019	0	297	❍	100%	◗
ATT&CK proc - tool	Sent.	❍	800	0	251	❍	100%	◗
ATT&CK TTP description	Sent.	❍	835	14	835	❍	100%	●
ATT&CK TTP detections	Sent.	❍	691	0	691	❍	100%	❍
ATT&CK TTP mitigation	Sent.	❍	1445	0	582	❍	100%	❍
CAPEC	Sent.	❍	273	0	189	❍	100%	◗
ATT&CK report	Doc.	2281	❍	0	797	❍	100%	◗
rcATT [15]	Doc.	1490	❍	12	215	❍	100%	●
APT&Cybercriminals [18]	Doc.	1676	❍	14	433	❍	9.70%	◗
WeliveSecurity [19]	Doc.	6090	❍	0	408	❍	2.40%	◗
apt notes [20]	Doc.	684	❍	14	336	❍	6.60%	◗
Malpedia [21]	Doc.	663	❍	14	522	❍	98.60%	❍

Legend: (Gran.) Granularity of labels. (T) Techniques. (ST) Sub-Techniques. (TA) Tactic, (NE) Negative Evidence highlights if there are sentences not mapped to TTPs, (Ratio) Ratio of Docs/sent containing TTP info, (Used) if the literature used it.

C. Supervision Availability, Scarcity, and Sparsity

A central finding of our ecosystem analysis is that the bottleneck for sentence-to-TTP learning is not the total amount of CTI text, but the scarcity and sparsity of *usable sentence-level supervision*. In particular, the limiting resource is the set of sentences that are explicitly mapped to one or more TTPs. This distinction matters because sentence-level corpora may contain many sentences, but only a minority carry technique labels. In our two primary sentence-level benchmarks, the proportion of TTP-positive sentences is substantially below the total corpus size (e.g., 58.19% in AnnoCTR and 21.22% in TRAM), underscoring the scarcity of high-signal annotated samples.

Consistent with this observation, the total number of TTP-positive sentence-level instances in benchmark datasets remains limited compared with what is potentially available from auxiliary sources. In contrast, MITRE-derived sentence-level resources and CTI report corpora provide substantially larger pools of supervision signal, albeit at different granularities and with different levels of direct usability for sentence-to-TTP training.

D. Potential Additional Supervision from CTI Reports

Many CTI corpora provide annotations at the document level, where each report is associated with a set of ATT&CK techniques. However, these annotations do not specify which individual sentences support each technique. For this reason, we do *not* treat document-level labels as sentence-level ground truth.

Instead, we estimate how much supervision these resources could potentially provide by defining *potential sentence-to-TTP supervision*. We define this quantity as the total number of document-level technique associations across reports. Intuitively, this value represents an upper bound on the amount of sentence-level supervision that could be recovered if each report-level technique were grounded to one or more evidence sentences.

This definition serves two purposes. First, it enables a meaningful comparison between document-level resources and sentence-level datasets from a supervision quantity perspective. Second, it motivates the retrieval-based method introduced in Section V, which aims to convert document-level CTI labels into sentence-level training pairs by grounding techniques to candidate evidence sentences within each report.

As shown in Table II, the amount of potential supervision available in document-level CTI corpora is substantial. This suggests that these resources represent an underutilized source of training signal for sentence-to-TTP learning.

1) Additional Supervision Sources: Our resource characterization includes several sources that, to the best of our knowledge, have not been used in prior CTI-to-TTP augmentation pipelines in this setting, namely MITRE ATT&CK *detections*, MITRE ATT&CK *mitigations*, and *Malpedia*.

TABLE II
CHARACTERIZATION OF PAPERS FOR SENTENCE TO TTP CLASSIFICATION USING WHICH DATASET

Dataset	#CTRs	#CTRs with Tec	#Potential Sent.	T/ST
ATT&CK report	2281	2040	2040	620
rcATT [15]	1490	1490	6342	215
APT & Cybercriminals [18]	1676	163	3384	433
WeliveSecurity [19]	6090	148	3170	408
apt notes [20]	684	45	1037	336
Malpedia [21]	663	654	8121	522

In particular, the MITRE-derived sources (*detections* and *mitigations*) account for a meaningful portion of the available MITRE sentence-level supervision pool, while *Malpedia* contributes a substantial share of the potential sentence-to-TTP supervision available in the document-level CTI collection. This suggests that considering resources beyond the datasets most commonly used in prior work may provide useful additional supervision for sentence-to-TTP learning.

E. Implications for Sentence-to-TTP Learning

Taken together, our analysis highlights three observations relevant to sentence-to-TTP learning. First, expert-annotated sentence-level supervision remains limited relative to the total volume of available CTI text. Second, the broader CTI-to-ATT&CK ecosystem contains additional supervision signals, but these are fragmented across annotation granularities and resource types. Third, benchmark-specific label restrictions can further reduce the amount of usable supervision, as annotations associated with out-of-scope techniques are typically discarded.

These observations suggest that improving sentence-to-TTP learning requires better utilization of the supervision already present across heterogeneous resources. In particular, two complementary strategies emerge: (i) *recovering* supervision that would otherwise be discarded under restricted label spaces through label-space alignment, and (ii) *transferring* document-level CTI annotations into sentence-level training pairs by grounding techniques to candidate evidence sentences. Section **??** presents a framework that operationalizes these two strategies.

V. METHODOLOGY: SUPERVISION UTILIZATION FRAMEWORK

Building on the findings from Section IV, we propose a supervision-utilization framework with two complementary augmentation pathways (Figure 1). First, we recover supervision from ATT&CK-derived resources that would otherwise be discarded under benchmark-specific

label constraints, by aligning mismatched labels through hierarchy- and semantic-based *recovery* (Section V-A). Second, we *transfer* document-level CTI supervision to the sentence level through retrieval-based grounding, producing evidence-linked sentence–TTP pairs from weakly labeled reports (Section V-B). Together, these two pathways expand usable sentence-level supervision from both structured ATT&CK knowledge and external CTI corpora. Export and allocation policies then control which recovered candidates are retained and how they are distributed across labels, balancing supervision quality, diversity, and class imbalance.

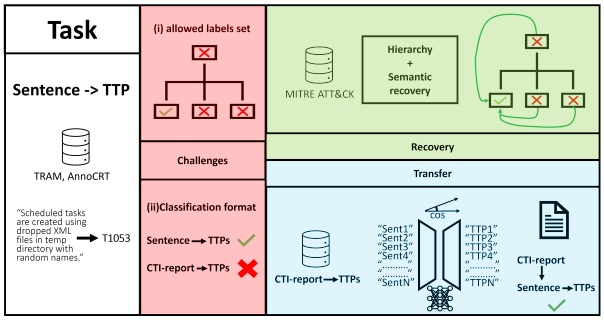

Fig. 1. Overview of the proposed framework for supervision recovery and transfer. The method combines label-space recovery and retrieval-based transfer to convert unused heterogeneous supervision sources into usable sentence-level training signals.

A. MITRE-Derived Augmentation: Recovery

MITRE ATT&CK provides multiple text fields that can be exploited as labeled supervision for technique learning, including technique descriptions and other procedure-like text associated with ATT&CK entries. These resources yield sentence-level supervision that can be added directly to the training set. However, under benchmark-specific label restrictions, a substantial fraction of this supervision would be discarded because many ATT&CK labels fall outside the admissible technique set. We therefore treat MITRE-derived augmentation as a controlled label-recovery process rather than as a simple data expansion step.

Given a benchmark-specific admissible label set, we apply two complementary recovery mechanisms to retain supervision that would otherwise be excluded: (i) *hierarchy-based recovery*, which exploits explicit parent–child relations in the ATT&CK ontology, and (ii) *semantic recovery*, which maps out-of-scope labels to semantically related admissible techniques when no suitable hierarchy-based mapping exists.

a) Hierarchy-based recovery: Not all sentence–TTP pairs extracted from ATT&CK can be used directly under benchmark constraints, because ATT&CK and the benchmark label space is a subset of the ATT&CK space. For example, a benchmark may include only a parent technique and none of its sub-techniques; rather than discarding the sub-technique pairs, we exploit the ATT&CK hierarchy to project labels back into the admissible benchmark label space and thus recover supervision that would otherwise be lost.

When a sentence is associated with a label outside the benchmark label space, we first check whether that label can be mapped to an in-scope parent technique. If so, we relabel the sentence with that parent and retain the pair. Conversely, if the benchmark includes only a single admissible sub-technique, we map its parent and sibling techniques to that sub-technique when the hierarchical relation makes the correspondence unambiguous. We apply this procedure only when the mapping is clear and unique, to avoid introducing ambiguous supervision. In this way, we preserve otherwise unusable training examples while remaining consistent with the benchmark label space.

b) Semantic recovery: Some ATT&CK labels cannot be recovered through the hierarchy because no admissible parent or sibling mapping is available. For these cases, we introduce a second augmentation path based on semantic similarity between ATT&CK techniques. We compute similarity between textual representations of techniques and map an out-of-scope label to its nearest admissible neighbor only when the similarity indicates meaningful conceptual alignment. This provides a controlled way to reuse otherwise discarded MITRE supervision beyond explicit ontology structure, complementing hierarchy-based recovery by expanding coverage where structural mapping alone is insufficient.

B. CTI-based augmentation: Transfer

Document-level CTI labels indicate which ATT&CK techniques occur in a report, but not which sentence expresses each technique. To transfer this weak supervision into sentence-level training data, we frame the sentence-to-TTP assignment as a text–text retrieval problem: given a CTI sentence, the goal is to identify the most semantically compatible ATT&CK technique description among the candidate techniques associated with the report. We operationalize this grounding step by introducing *TTP-SiamAlign*: a siamese bi-encoder, which embeds CTI sentences and ATT&CK technique texts into a shared representation space and ranks candidate techniques by cosine similarity.

The architecture is *siamese* [22] because both inputs are processed by the same Transformer encoder with tied weights, ensuring a consistent embedding space across CTI-sentences and TTP-descriptions. It is a *bi-encoder* [23] because CTI-sentences and TTPs representations are computed independently, which enables efficient retrieval by precomputing embeddings for all candidate techniques once and reusing them across queries. For

each input text, we obtain a single vector by mean pooling the token-level hidden states while masking padding tokens, followed by ℓ_2 normalization. Candidate techniques are then ranked by cosine similarity between the normalized sentence and technique embeddings.

We train *TTP-SiamAlign* on available sentence-level sentence–TTP pairs, including benchmark supervision and aligned MITRE-derived pairs, using contrastive learning with in-batch negatives. Within each mini-batch, the gold sentence–technique pair is treated as the positive match, while other technique descriptions in the batch act as negatives. This objective encourages the model to assign higher similarity to true sentence–technique pairs than to mismatched alternatives, yielding a retrieval function that supports both semantic label recovery and pseudo-label mining from weakly labeled CTI reports. The graphical overview of *TTP-SiamAlign* is shown in Figure 3 in the Appendix.

1) Transfer from Document-Labeled CTI.: Document-labeled CTI corpora provide report-level technique sets but no sentence-level grounding. We therefore use report labels as *weak constraints*: each report is segmented into sentences, embedded with *TTP-SiamAlign*, and each sentence is scored only against the techniques assigned to that report after label-space recovery. This constrained retrieval reduces false positives and avoids searching the full ATT&CK space for every sentence. The resulting scored sentence–technique pairs are exported as pseudo-labels using confidence controls such as thresholding.

VI. Experimental Setup

Our experimental design aims to isolate two complementary questions: (i) the effect of *Recovery*, i.e., MITRE-derived augmentation and label-space recovery, and (ii) the effect of *Transfer*, i.e., weak-to-strong transfer from document-labeled CTI through retrieval-based mining. To study these factors clearly, we divide the evaluation into two parts. First, we conduct an *intrinsic evaluation* to assess the quality of the augmentation mechanisms themselves, including the reliability through expert-validated augmentation candidates and the effectiveness of the proposed method *TTP-SiamAlign* used for retrieval-based mining. Second, we perform a *downstream evaluation* to measure whether these augmentation strategies improve the target task of sentence-to-TTP classification.

A. Intrinsic Evaluation

Our intrinsic evaluation assesses the quality of the augmentation process itself, independent of its downstream effect on sentence-to-TTP classification. We consider two complementary aspects. First, for both MITRE-derived (recovery) and CTI-derived augmentation (transfer), we evaluate the plausibility of the generated sentence–TTP pairs through manual expert assessment. In addition, for the retrieval-based Transfer setting, we evaluate *TTP-SiamAlign* as a mining backbone using retrieval effectiveness, including Hit@1, to determine how accurately relevant TTP labels are recovered from weakly labeled document-level CTI data. We evaluate our method as a ranking task using metrics common in link prediction and information retrieval [24], since the goal is to identify the most likely TTP for a given sentence. We report Hit@1, which measures whether the correct technique is ranked first. This is well aligned with our setting because downstream classification relies on these ranked sentence–TTP matches to generate augmented supervision.

B. Downstream Evaluation

The downstream evaluation measures the impact of augmentation on the target task of sentence-to-TTP classification. We report classification performance on TRAM (50 labels) and AnnoCTR (111 labels), which serve as the main benchmarks throughout the paper. To establish a reliable point of comparison, we first quantify baseline performance without augmentation. We then measure the effect of augmentation on classification performance, expressed primarily in terms of F1-score. In addition, we examine the impact of preprocessing, as this is a commonly used but inconsistently reported component in prior work and may substantially affect downstream performance, as detailed in Appendix XI-E.

1) Baselines: Following the systematization provided by [2], we adopt TRAM and AnnoCTR as benchmark datasets for evaluation. These datasets represent the most suitable publicly available corpora for sentence-level TTP extraction under the MITRE ATT&CK ontology.

We summarize and implement the two dominant families of approaches, as also highlighted by [25]: *(i) classification-based methods* and *(ii) generative LLM-based methods*.

a) Classification-based approaches.: These methods treat TTP extraction as a multi-label text classification task. Pre-trained encoder models (e.g., BERT and its CTI-specific variants) are fine-tuned on labeled training data derived from TRAM or AnnoCTR. The encoder generates contextual embeddings that are then used for classification.

b) Generative approaches.: Generative methods leverage LLMs to predict TTP identifiers via instruction-following prompts. Depending on the configuration, such as zero-shot, incorporate retrieval-augmented generation (RAG) using ATT&CK descriptions as external context, or be adapted via supervised fine-tuning (SFT) on instruction-formatted training data. Instead of directly

optimizing classification logits, these approaches generate textual outputs from which TTP identifiers are extracted.

2) Training Configuration: All classification baselines are trained under a unified multi-label fine-tuning setup: AdamW optimization, BCEWithLogitsLoss, linear learning-rate decay, gradient clipping, fixed random seeds, and early stopping on validation loss. Where applicable, class imbalance is addressed through positive-label weighting. Generative baselines, when fine-tuned, are trained using supervised fine-tuning with LoRA adapters on an instruction-tuned LLM. Optimization follows a causal language modeling objective under 16-bit precision. Inference outputs are parsed to recover predicted technique IDs. A detailed overview of hyperparameters and optimization settings is provided in Table XII in the Appendix. To preserve a clean evaluation protocol, all augmented instances are added *exclusively* to the training split and are never used in the construction of the validation or test sets. For CTI-derived augmentation, we further control for redundancy by removing duplicate documents across corpora.

3) Export and Allocation Strategies for Augmentation: Once the two augmentation methods (described in Sections V-A - V-B) have produced candidate sentence–TTP pairs, the next question is how to incorporate them into the downstream task. This is non-trivial because the TTP label distribution is highly imbalanced: some classes are associated with substantially more examples than others. Naively adding all augmented candidates may therefore further reinforce already dominant labels, while overly restrictive selection may discard useful variability and limit the benefits of augmentation.

To make augmentation controllable, we decompose it into two explicit decisions: *export* and *allocation*. *Export* determines which candidate pairs are retained from the recovered or mined pool, for example, by selecting the highest-confidence pairs. *Allocation* determines how many exported pairs are added per label.

We study four allocation strategies. *Global* adds all exported pairs without label-wise constraints, maximizing supervision volume but also preserving the original skew. *Uniform* distributes additions evenly across labels, encouraging balanced exposure and preventing head labels from dominating the augmented set. *Cap-by-max* limits the number of added pairs for each label relative to the largest class, reducing uncontrolled growth while still allowing augmentation for all labels. Finally, *Performance-driven* allocates more augmentation to labels or configurations that appear to benefit most under validation feedback, aiming to use the budget where additional supervision is most effective.

This separation between export and allocation allows us to analyze augmentation not simply as "more data," but as a controlled design space over candidate quality, diversity, and class balance. Full implementation details and exact selection rules for each strategy are provided in Appendix XI-C.

VII. Intrinsic Evaluation

We evaluate the two augmentation strategies by first providing a human evaluation over a stratified extracted sample of 50 labels per augmentation. Then we evaluate *TTP-SiamAlign* under a retrieval-based sentence–TTP setting and compare it against alternative learning paradigms that exploit different subsets of the available supervision.

A. Manual evaluation

We manually assessed the quality of the generated supervision through an expert-based plausibility study. For each augmentation source, we first expanded the data into sentence–label pairs, removed exact duplicate pairs, and retained only labels belonging to the benchmark-allowed label set. We then drew a stratified random sample to avoid over-representing frequent techniques, yielding 50 hierarchy-recovered pairs, 50 semantic-recovered pairs, and 50 CTI pseudo-labeled sentence–TTP pairs. Two experts with ATT&CK familiarity independently annotated each sampled instance. For cases of disagreement, a third expert reviewed the instance, and the final label was determined by majority vote over the two judgments. For recovered pairs, annotators were shown the sentence and the recovered admissible label, and judged whether the recovered label constituted a plausible weak supervision target for the sentence, as already proposed by [4]. This evaluation is intended to assess whether the generated supervision is reliable enough to serve as a weak training signal, rather than to measure exact label correctness against a gold standard, which is unavailable.

TABLE III
MANUAL EXPERT EVALUATION OF GENERATED SUPERVISION QUALITY. FOR EACH SOURCE, TWO EXPERTS ASSESSED WHETHER THE GENERATED SENTENCE–TTP PAIR WAS *plausible*, *mild*, OR *negative* AS WEAK SUPERVISION. WE REPORT THE FINAL AGREED LABELS AFTER DISCUSSION.

Source	n	Plausible	Mild	Negative
Hierarchy	50	X (X%)	X (X%)	X (X%)
Semantic	50	X (X%)	X (X%)	X (X%)
CTI-retrieved	50	X (X%)	X (X%)	X (X%)
Overall	150	X (X%)	X (X%)	X (X%)

Placeholder for evaluation result and table explanation.

B. TTP-SiamAlign *intrinsic evaluation*

We consider the setting in which a CTI report is labeled only at the document level with a set of ATT&CK techniques, without sentence-level grounding. The goal in this scenario is to recover sentence–TTP associations, i.e., to identify for each reported TTP the sentence in the report that supports it.

Although the target CTI reports provide only document-level labels, different methods can exploit additional supervision sources during training or inference. In Table IV, *Sent–TTP* denotes access to external sentence-level supervision in the form of annotated sentence–TTP pairs; *CTI* denotes access to document-labeled CTI reports, where each report provides a set of TTPs without sentence-level alignment; this signal constrains retrieval and requires the model to produce a complete sentence–TTP mapping for the TTPs listed in the report. *TTP descr.* denotes access to the natural-language descriptions of ATT&CK techniques, which provide semantic grounding for the label space. Different paradigms exploit different subsets of these sources, resulting in different trade-offs between direct supervision, semantic grounding, and robustness as the label space grows.

We therefore compare *TTP-SiamAlign* against two representative baselines. *Unsupervised retrieval* matches CTI sentences to ATT&CK technique descriptions using pretrained semantic representations without task-specific training, testing whether semantic similarity alone is sufficient for sentence–TTP matching in document-labeled CTI reports. *Supervised classification* is trained on external labeled sentence–TTP pairs and benefits from direct supervision, but treats TTPs as discrete labels rather than semantically meaningful text. This reflects the dominant formulation in prior sentence-level TTP classification work. In contrast, *TTP-SiamAlign* leverages all three sources listed in Table IV: external *Sent–TTP* supervision, document-level *CTI* constraints, and *TTP descr.*. By embedding CTI sentences and ATT&CK technique descriptions into a shared representation space, it combines sentence-level supervision with semantic access to the label space and naturally supports constrained retrieval from document-labeled CTI corpora.

For a fair comparison, all methods are evaluated on the same held-out test sets. Training and validation splits are used only for the supervised classifier and *TTP-SiamAlign*, whereas unsupervised retrieval requires no task-specific training. The evaluation spans increasingly challenging label spaces: TRAM with 50 labels, AnnoCTR with 111 labels, and *ALL* with 480 labels. The *ALL* setting combines all available sentence–TTP pairs from both benchmark datasets and MITRE-derived sources, including ATT&CK descriptions, procedures, etc.

TABLE IV
CHARACTERIZATION OF LEARNING PARADIGMS FOR SENTENCE-TO-TTP EXTRACTION ON THE VALIDATION SET

Method	**Sources**			**hit@1**		
	Sent–TTP	CTI	TTP descr	TRAM	Anno CTR	ALL
Unsup	×	✓	✓	0.07	0.04	0.05
Sup	✓	×	×	0.87	**0.82**	0.65
Siamese	✓	✓	✓	**0.89**	0.80	**0.70**

Table IV shows that *TTP-SiamAlign* is the most appropriate backbone, although not because it uniformly dominates the supervised baseline in every setting. In smaller label spaces, its performance is comparable to supervised classification. On TRAM, the *TTP-SiamAlign* model slightly outperforms the supervised classifier, whereas on AnnoCTR the supervised classifier remains marginally stronger. However, as the label space becomes larger and more heterogeneous, the advantage shifts toward *TTP-SiamAlign* formulation: in the *ALL* setting, it achieves the best result, suggesting better scalability when semantic distinctions across techniques become finer-grained and harder to separate. This likely occurs because, in larger label spaces, standard supervised approaches treat labels as arbitrary class IDs and therefore do not exploit semantic relationships between techniques. *TTP-SiamAlign*, instead, encodes the label space itself through technique-description embeddings, which helps preserve finer-grained distinctions.

We therefore adopt *TTP-SiamAlign* not because it consistently outperforms supervised classifiers in closed-label prediction, but because it offers the best trade-off for our setting. It remains competitive when the label space is limited, scales more effectively as label diversity increases, and can produce sentence–TTP matches for all techniques associated with a CTI report, including those outside the scope of supervised training. Most importantly, unlike standard supervised classification, it can be directly applied to the document-labeled CTI setting by combining *CTI* constraints with *TTP descr.*, making it the most suitable backbone for the pseudo-label mining pipeline developed in the remainder of the paper.

VIII. DOWNSTREAM EVALUATION

We evaluate two augmentation settings: **(i)** *Recovery*, which augments training data with MITRE-derived supervision and label-space recovery, and **(ii)** *Transfer*, which transfers document-labeled CTI reports to sentence-ttp pairs through retrieval-based sentence mining. Unless stated otherwise, all results are reported as the mean over 10 independent runs.

A. Baseline Performance

Before analyzing augmentation, we establish strong baselines for the two dominant families of sentence-to-TTP extraction methods identified in prior systematization work: *classification-based* approaches and *generative* LLM-based approaches. Table V summarizes the strongest classification and generative baselines on TRAM and AnnoCTR. We use the best reference model throughout the remainder of the paper to assess whether augmentation yields gains beyond those already achieved by the best model.

a) Classification baselines.: Classification-based methods formulate sentence-to-TTP extraction as a multi-label text classification task over the benchmark label space, which remains the dominant supervised setting in prior work. Consistent with this formulation, we evaluate 13 encoder models, spanning both general-purpose and security-specialized representations, under the unified multi-label fine-tuning protocol described in Section VI. Full results are reported in Table XV in the Appendix

b) Generative baselines.: Generative approaches predict TTP identifiers through instruction-following generation rather than direct classification. We therefore evaluate Llama 3 8B-Instruct in the supervised fine-tuning setting, which was identified in [2] as the strongest among six evaluated configurations, including zero-shot prompting, few-shot prompting, and retrieval-augmented generation. Unlike classification, where performance differences across models are relatively limited and motivate broader model comparisons, the generative setting exhibits a clear best choice, making supervised fine-tuning the most appropriate baseline.

TABLE V
BEST MODELS FOR SENTENCE-TO-TTP CLASSIFICATION

Best Models	**TRAM** [17]			**AnnoCTR** [16]		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Classification	**77.20**	62.80	**66.57**	**70.49**	51.03	**56.62**
Generation	69.05	**64.19**	64.16	52.13	**54.10**	47.87

B. Effect of MITRE-Derived Augmentation

Prior work has already considered MITRE as an augmentation source for sentence–TTP classification. Our focus here is not on whether MITRE-derived supervision can help in principle, but on two questions that remain underexplored: how much of this supervision should be injected, and how supervision associated with labels outside the benchmark label set should be handled.

1) MITRE-only augmentation: To address the first question, Table XVI in the Appendix evaluates the effect of MITRE-derived sentence–TTP pairs under different augmentation budgets and allocation policies. The main finding is that additional MITRE supervision is beneficial only when its injection is controlled. Simply appending all available MITRE pairs does not yield consistent gains, whereas budgeted allocation strategies lead to more reliable improvements by distributing added supervision more deliberately across the label space. This result is important because it shows that the benefit does not stem from adding more data per se, but from adding supervision in a way that better matches the needs of the benchmark label distribution. The best augmentation for this source is reported in the second row of Table VI (✓ × ××).

2) Recovery: We evaluate the *recovery* mechanisms introduced in Section V-A. The corresponding results are reported in Table VI: row four for Hierarchy-based augmentation (✓✓ × ×) and row five for Hierarchy+Embedding augmentation (✓✓✓×). Both recovery strategies are evaluated on top of *MITRE-only augmentation*, as expanding the MITRE hierarchy is only meaningful when MITRE-derived samples are included.

a) Hierarchy augmentation.: Hierarchy-based recovery addresses a major source of supervision loss: MITRE sentences whose original labels fall outside the benchmark label set. Instead of discarding these cases, hierarchy recovery maps them to admissible parent/peer/child techniques when such a mapping is unambiguous. The results show that this recovery step is highly beneficial, especially on ANNOCTR, where it yields the strongest overall gains across all the other augmentation types. This indicates that a substantial portion of the improvement comes not from introducing entirely new information, but from recovering useful supervision that would otherwise be excluded due to label-space mismatch.

b) Hierarchy + Embedding augmentation.: Hierarchy-based recovery alone cannot resolve all out-of-set labels, since some techniques do not have a usable parent-based mapping. Embedding-based recovery broadens coverage by semantically aligning such cases to admissible techniques, increasing the amount of supervision that can be recovered beyond what the ATT&CK hierarchy alone supports. The results suggest that this additional recovered supervision is complementary to hierarchy-based augmentation on TRAM. On AnnoCTR, the combination is slightly weaker than hierarchy recovery alone, but still outperforms MITRE-derived augmentation without recovery.

C. Effect of CTI-Derived Augmentation: Transfer

Pseudo-label mining is performed by leveraging *TTP-SiamAlign* described in Section V-B. For each report, sentences are scored only against the techniques as-

TABLE VI

ABLATION STUDY OF AUGMENTATION SOURCES, RECOVERY MECHANISMS, AND ALLOCATION STRATEGIES ON TRAM AND ANNOCTR. COLUMNS (1)–(4) INDICATE WHETHER MITRE-DERIVED AUGMENTATION (M), HIERARCHY-BASED RECOVERY (H), EMBEDDING-BASED RECOVERY (E), AND CTI-DERIVED AUGMENTATION (C) ARE ENABLED. COLUMNS G AND U DENOTE GLOBAL AND UNIFORM ALLOCATION, RESPECTIVELY. AN ASTERISK (*) MARKS THE BEST-PERFORMING CONFIGURATION FOR EACH DATASET, WHICH INCLUDES E FOR TRAM AND EXCLUDES IT FOR ANNOCTR.

(1)	(2)	(3)	(4)	Augm		**TRAM** [17]				**AnnoCTR** [16]			
M	**H**	**E**	**C**	**G**	**U**	**Train pairs**	**Precision**	**Recall**	**F1**	**Train pairs**	**Prec.**	**Rec.**	**F1**
×	×	×	×	×	×	12017	77.20	62.80	66.57	2934	70.49	51.03	56.62
✓	×	×	×	×	✓	+6012 (50.03%)	74.43	67.37	68.52	+3525 (120.14%)	69.69	55.38	59.42
				✓	×	+7929 (65.98%)	75.26	61.33	65.63	+6535 (222.73%)	68.89	59.88	61.58
×	×	×	✓	×	✓	+5127 (42.61%)	70.18	68.85	67.64	+523 (17.83%)	71.09	55.72	59.92
				✓	×	+2315 (19.24%)	**77.67**	59.65	64.71	+2289 (78.02%)	**74.83**	51.00	57.89
✓	✓	×	×	×	✓	+12 217 (101.66%)	72.62	64.90	66.47	+16 011 (545.71%)	68.77	**63.42**	**63.46**
				✓	×	+9476 (78.85%)	73.44	67.03	67.73	+7276 (247.99%)	65.50	58.51	59.50
✓	✓	✓	×	✓	×	+12 435 (103.48%)	72.69	68.99	68.81	+16 163 (550.89%)	68.56	60.00	61.65
				×	✓	+9694 (80.67%)	71.99	66.64	67.28	+7365 (251.02%)	68.93	57.65	60.26
✓	×	×	✓	✓	×	+25 719 (214.02%)	75.35	65.62	68.04	+22 419 (764.11%)	74.65	49.94	57.13
				×	✓	+11 930 (99.28%)	71.61	68.78	68.30	+4046 (137.90%)	68.02	57.30	60.23
✓	✓	*	✓	✓	×	+30 225 (251.52%)	73.26	65.09	67.05	+31 895 (1087.08%)	73.66	52.04	57.38
				×	✓	+15 823 (131.67%)	69.17	**73.26**	**69.14**	+7828 (266.80%)	65.57	59.97	60.70

sociated with that document. Exported pseudo-labeled sentence–TTP pairs are then selected and added to the training set.

As the only augmentation strategy that introduces in-distribution samples, CTI-derived augmentation is also the only one that improves classification precision (third row of Table **??**). Although its absolute F1 gains are smaller than those of MITRE-derived augmentation, which mainly increases recall, it is the most sample-efficient method among the augmentation and budgeting strategies compared in Appendix XI-G. Concretely, adding only 2.12% and 17.83% more training pairs produces F1 gains of 1.59% on TRAM and 5.83% on AnnoCTR, respectively.

D. Ablation Study

To isolate which design choices drive the observed gains, we perform controlled ablations over (i) augmentation sources and (ii) allocation controls. Specifically, we vary the inclusion of MITRE-derived pairs (M), label-space recovery mechanisms (hierarchy-based mapping, H; and embedding-based mapping, E), and CTI-derived pseudo-labels (C), while toggling over the best allocation policies (global vs. uniform budgeting; G/U). Table **??** reports, for each configuration, the resulting increase in training pairs and the downstream performances

Overall, the best TRAM result is achieved by combining all three supervision sources with a *uniform* allocation policy (*M+H/E+C*), suggesting that the proposed *Recovery* and *Transfer* augmentations provide complementary benefits for downstream sentence-to-TTP classification. On AnnoCTR, however, the strongest configuration is MITRE-derived augmentation with hierarchy-based recovery alone (*M+H*); adding CTI-derived pseudo-labels improves precision but causes a substantial reduction in recall. This contrast indicates that augmentation quality is governed not merely by the number of added pairs, but by how supervision source (MITRE vs. CTI), label alignment (H/E), and allocation control (G/U) interact under different label-space conditions.

E. Comparison with State-of-the-Art

We compare our supervision-utilization framework with four augmentation approaches discussed in Section III: TTPXHunter [12], EDA+BT [11], HMCAT [13], and the procedure-based augmentation of You *et al.* [14]. These baselines span the main augmentation paradigms proposed for sentence-to-TTP extraction, including lexical perturbation (EDA+BT), context-preserving token substitution (TTPXHunter), LLM-based text generation (HMCAT), and procedure-driven supervision from ATT&CK documentation (You *et al.*). Illustrative examples of the augmentation methods are provided in Table XI-H in the Appendix.

Table VII summarizes the resulting performance on TRAM and AnnoCTR. Overall, our approach consistently achieves the strongest recall and the best F1 scores across both datasets. On TRAM, our method outperforms

the best prior augmentation approach by $\sim$2% while also providing the largest recall improvement ($\sim$10%). Although some baselines achieve slightly higher precision, their lower recall limits the overall F1 performance. These results indicate that our supervision-augmentation strategy primarily improves the model's ability to detect a broader set of techniques without sacrificing overall classification quality.

TABLE VII
COMPARISON OF "OUR" APPROACH WITH PREVIOUS WORK

Augment method	TRAM [17]				AnnoCTR [16]			
	δ +%	Prec.	Rec.	F1	δ +%	Prec.	Rec.	F1
Baseline	-	**77.20**	62.80	66.57	-	70.49	51.03	56.62
TTPXHunter	+49	75.05	66.30	68.07	+133	**72.40**	49.80	56.30
EDA+BT	+40	76.52	60.28	65.43	+73	68.56	47.01	53.19
HMCAT	+24	78.71	63.31	67.04	+32	70.06	51.30	56.39
You et al.	+59	74.20	64.20	66.99	+186	68.67	53.47	57.36
Our	+132	69.17	**73.26**	**69.14**	+546	68.77	**63.42**	**63.46**

The improvements are even more pronounced on AnnoCTR. Our method increases F1 by more than 12%, representing a substantially larger improvement than competing augmentation approaches, because prior methods provide at most $\sim$1% improvement.

A key distinction between our method and previous augmentation strategies lies in the nature of the generated supervision. Prior approaches mainly rely on token/word-level transformation or synthetic generation of new sentences. While these methods increase dataset size, they do not address the fundamental supervision bottleneck caused by the limited amount of supervision available. In contrast, our approach expands the supervision pool by (i) extracting previously unused supervision from the ATT&CK knowledge base, (ii) recovering supervision through hierarchy- and embedding-based label alignment, and (iii) transferring document-level CTI labels into sentence-level training signal through retrieval-based grounding. As a result, the augmented data remains semantically aligned with the ATT&CK ontology, leading to more informative supervision and improved downstream performance.

IX. DISCUSSION

A. Key Findings and Practical Implications

Across all experiments, the outcomes point to the same conclusion: our combined augmentation method is beneficial to the sentence-ttp classification

a) Recovery: A first key finding is that a substantial fraction of potentially useful ATT&CK supervision would remain excluded without recovery. Hierarchy-based recovery and embedding-based recovery both expand the usable augmentation pool by mapping otherwise discarded labels back into the benchmark label space. The practical implication is that augmentation quality depends not only on the source of additional data, but also on whether label-space mismatch is explicitly resolved. Structural recovery captures cases where supervision can be transferred through the ATT&CK hierarchy, while semantic recovery extends coverage to cases where no direct benchmark label is available but a close in-scope proxy exists. Together, these mechanisms make augmentation more complete and better aligned with the downstream task.

b) Transfer: In contrast to MITRE-derived augmentation, document-level CTI mining introduces supervision that is closer to the target data distribution. This makes it particularly valuable as a complementary source of contextual variability, and explains why it is the only augmentation strategy that consistently improves precision. At the same time, pseudo-label transfer is inherently noisier than ATT&CK-derived augmentation, so unrestricted injection can reduce the benefit by introducing incorrect or weakly grounded sentence–TTP pairs. Our results therefore, show that CTI-derived augmentation is most effective when combined with specific export and allocation controls. In practice, weak supervision from CTI should be treated as a high-value but potentially high-risk resource: useful for improving realism and class discrimination, but only when confidence filtering and budgeting prevent noise from dominating the added signal.

c) Main takeaway: ATT&CK-derived augmentation is most useful for improving coverage, especially in low-resource settings, while CTI-derived transfer contributes complementary in-distribution evidence that can sharpen precision.

B. Model improvement vs Data improvement

Our results highlight an important challenge in CTI-to-ATT&CK sentence classification: the bottleneck derived from models is comparable to supervision data availability and alignment. Existing work often focuses on architectural improvements or surface-level text augmentation. However, our experiments show that leveraging structured knowledge sources and weak supervision can produce larger gains than traditional augmentation techniques.

1) Model improvement.: To quantify the effect of encoder choice, we examine the spread of baseline performance in Table XV. Most pretrained encoders lie within a relatively narrow performance band, with the best model improving over the median by only $+7.9\%$ on TRAM and $+10.5\%$ on AnnoCTR. This suggests that

encoder selection has a moderate impact, reinforcing the need to focus on supervision quality and augmentation design rather than model choice alone.

2) Data improvement.: Augmentation yields gains of similar or greater magnitude than encoder selection. Relative to the strongest baseline backbone, it improves F1 by $+3.9\%$ on TRAM and $+12.1\%$ on AnnoCTR (Table VII). These improvements are driven mainly by recall, suggesting that augmentation primarily expands coverage of technique expressions rather than improving precision.

Taken together, these results show that improvements obtained through supervision expansion are comparable in magnitude to those obtained through model selection. While different pretrained encoders produce measurable performance differences, expanding the supervision signal through our proposed augmentation can yield improvements of a similar scale, particularly on datasets with smaller training sets such as AnnoCTR. These results suggest that supervision utilization is comparable to model selection (complementary rather than substitutive), so improvements should be pursued jointly.

C. Per-Technique Impact of Augmentation

To analyze which ATT&CK techniques benefit most from augmentation, we compare per-label performance between the best baseline and the best augmented model, and visualize the relationship between label support and F1 improvement in Figure X. To avoid instability from extremely rare labels, we restrict the analysis to techniques with at least three test instances.

Across both TRAM and AnnoCTR, a clear and consistent pattern emerges: the largest improvements are concentrated among techniques with low to moderate support. As shown in Figure 2, techniques with support in the lower range exhibit the largest positive gains in ΔF1, with several techniques achieving improvements above +0.6. In contrast, techniques with higher support cluster tightly around zero, indicating limited sensitivity to additional supervision.

This distribution indicates that augmentation primarily benefits underrepresented techniques. For these labels, the dominant effect is a reduction in false negatives, leading to substantial recall gains and, consequently, large increases in F1. Importantly, these improvements are systematic rather than isolated, affecting multiple low-support techniques across both datasets. On the other hand, as label support increases, the spread of ΔF1 narrows considerably, clustering close to zero on both datasets, indicating that augmentation has only a limited impact on well-represented techniques.

D. Embedding Geometry

To better understand how augmentation reshapes the representation space, we analyze four embedding-

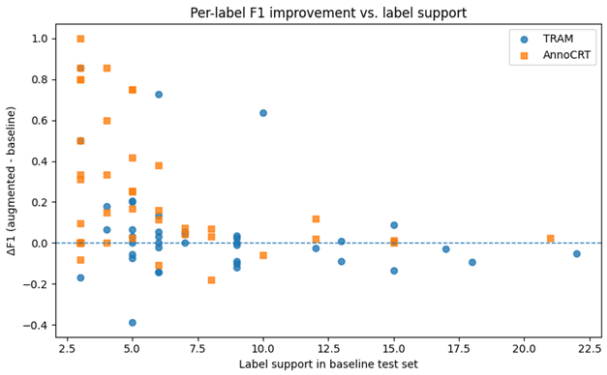

Fig. 2. Per-label ΔF1 (augmented baseline) as a function of baseline label support for TRAM and AnnoCTR. Improvements are predominantly observed for low-support techniques, particularly in AnnoCTR, while performance differences converge toward zero as support increases, indicating limited impact of augmentation on well-represented labels.

geometry diagnostics computed from sentence embeddings. *Intra-centroid distance* measures the average distance between samples and the centroid of their assigned technique class, with lower values indicating stronger within-class cohesion. *Inter-centroid distance* measures the average distance between class centroids, with higher values indicating better global separation. The *silhouette score* captures the balance between cohesion and separation, with higher values indicating clearer cluster structure. Finally, *kNN purity@10* measures the fraction of the ten nearest neighbors that share the same label, reflecting local label consistency.

Table VIII reports the relative change of each metric with respect to the baseline representation of the corresponding dataset. Across both TRAM and AnnoCTR, the clearest and most consistent effect of augmentation is the improvement in *silhouette* and *kNN purity*, indicating better local organization of the embedding space. On TRAM, MITRE-derived augmentation gives the strongest gains, improving silhouette by +85.1% and kNN purity by +90.8%. CTI-derived augmentation also improves local structure, but more moderately. On AnnoCTR, the same trend holds for silhouette, while the largest kNN gains are obtained by MITRE+H+E. By contrast, improvements in cluster separation are less uniform, and intra-centroid distance never substantially decreases relative to the baseline. This suggests that augmentation primarily improves the boundaries and local neighborhood structure of the embedding space, rather than increasing within-cluster cohesion.

A second consistent pattern is that improvements in local structure do not necessarily translate into improved *global* class separation. In particular, inter-centroid distance decreases for all settings on AnnoCTR and for several of them on TRAM. This suggests that augmen-

TABLE VIII
RELATIVE CHANGE (%) IN EMBEDDING-GEOMETRY DIAGNOSTICS AFTER AUGMENTATION RELATIVE TO THE BASELINE OF EACH DATASET. LOWER IS BETTER FOR INTRA-CENTROID DISTANCE (INTRA), WHEREAS HIGHER IS BETTER FOR INTER-CENTROID DISTANCE (INTER), SILHOUETTE (SILH.), AND KNN PURITY@10 (KNN). * MARKS THE BEST-PERFORMING CONFIGURATION.

				intra(%)(↓)		inter(%)(↑)		silh.(%)(↑)		kNN(%)(↑)	
M	**H**	**E**	**C**	**TRAM**	**ANNO**	**TRAM**	**ANNO**	**TRAM**	**ANNO**	**TRAM**	**ANNO**
✓	×	×	×	+15.1	+15.6	**+10.0**	-63.0	**+85.1**	+69.7	**+90.8**	+27.6
×	×	×	✓	**-0.2**	**+0.8**	-9.8	**-52.2**	+55.6	+50.4	+29.6	-10.3
✓	✓	*	×	+12.5	+18.2	+1.2	-64.9	+69.7	+69.8	+87.8	**+36.5**
✓	×	×	✓	+7.6	+15.4	-2.3	-64.2	+71.3	**+75.0**	+77.0	+26.0
✓	✓	*	✓	+7.0	+17.9	-7.7	-65.4	+73.0	+67.8	+77.9	+35.9

tation often makes neighborhoods more label-consistent without uniformly pushing class centroids farther apart. In other words, the main benefit appears to come from improving *local cohesion and neighborhood purity*, rather than from producing a more globally separated embedding space.

Overall, the geometry diagnostics suggest that augmentation mainly improves the *local* organization of the embedding space by making neighborhood structure more label-consistent, while its effect on *global* centroid separation remains mixed and can even decrease.

E. Limitations and Future Work

Despite the improvements demonstrated in this work, several limitations remain. First, the evaluation is limited to two benchmark datasets, TRAM and AnnoCTR. Although these datasets are widely used in CTI-to-ATT&CK research, additional evaluation on larger and more diverse CTI corpora would further validate the generality of the proposed approach.

Second, the current augmentation pipeline treats supervision sources independently. A promising direction for future research is to develop adaptive augmentation policies that dynamically balance structured knowledge, weak supervision, and synthetic data generation based on dataset characteristics.

Finally, future work could also revisit the evaluation protocols used for CTI-to-ATT&CK sentence classification. This work is not intended as a benchmarking study; instead, we adopt the evaluation metrics, datasets, and experimental setup proposed in [2] to ensure comparability with prior work. Our primary objective is to study how data augmentation and supervision expansion influence model performance under the current state-of-the-art evaluation framework. Future work could explore different metrics other than weighted average and different granularities other than senntence-ttp.

X. CONCLUSION

In this work, we studied the supervision bottleneck in CTI-to-ATT&CK sentence-to-TTP classification. Rather than focusing on new model architectures, we investigated how existing supervision sources can be more effectively utilized and expanded. We proposed a supervision-expansion framework that (i) extracts additional training pairs from MITRE ATT&CK knowledge sources, (ii) recovers supervision through hierarchy- and embedding-based label-space alignment, and (iii) transfers document-level CTI labels into sentence-level training signal via retrieval-based mining.

Across two widely used benchmarks (TRAM and AnnoCTR), our approach consistently improves recall and achieves the best overall F1 scores compared to existing augmentation strategies. The results show that structured knowledge from ATT&CK provides reliable supervision anchors, while CTI-derived pseudo-labels contribute complementary linguistic diversity when applied under controlled allocation policies.

Our findings highlight that the main limitation in CTI-to-ATT&CK extraction is not model capacity but supervision availability and alignment. By systematically expanding and aligning supervision sources, it is possible to significantly improve sentence-level technique classification without relying on more complex architectures. We hope this work encourages future research on supervision-centric approaches for CTI analysis and other structured threat intelligence tasks.

REFERENCES

[1] MITRE, "ATT&CK," https://attack.mitre.org/, 2023.

[2] M. Büchel, T. Paladini, S. Longari, M. Carminati, S. Zanero, H. Binyamini, G. Engelberg, D. Klein, G. Guizzardi, M. Caselli *et al.*, "{SoK}: Automated {TTP} extraction from {CTI} reports–are we there yet?" in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 4621–4641.

[3] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, "Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources," in *Proceedings of the 33rd annual computer security applications conference*, 2017, pp. 103–115.

[4] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi, "Ctibench: A benchmark for evaluating llms in cyber threat intelligence," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50 805–50 825, 2024.

[5] H. Zhang, G. Shen, C. Guo, Y. Cui, and C. Jiang, "Ex-action: Automatically extracting threat actions from cyber threat intelligence report based on multimodal learning," *Security and Communication Networks*, vol. 2021, no. 1, p. 5586335, 2021.

[6] N. Rani, B. Saha, V. Maurya, and S. K. Shukla, "Ttphunter: Automated extraction of actionable intelligence as ttps from narrative threat reports," in *Proceedings of the 2023 australasian computer science week*, 2023, pp. 126–134.

[7] F. I. Rahman, S. M. Halim, A. Singhal, and L. Khan, "Alert: A framework for efficient extraction of attack techniques from cyber threat intelligence reports using active learning," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2024, pp. 203–220.

[8] V. Orbinato, M. Barbaraci, R. Natella, and D. Cotroneo, "Automatic mapping of unstructured cyber threat intelligence: An experimental study:(practical experience report)," in *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2022, pp. 181–192.

[9] S. Della Penna, R. Natella, V. Orbinato, L. Parracino, and L. Pianese, "Cti-hal: A human-annotated dataset for cyber threat intelligence analysis," *arXiv preprint arXiv:2504.05866*, 2025.

[10] U. Kumarasinghe, A. Lekssays, H. T. Sencar, S. Boughorbel, C. Elvitigala, and P. Nakov, "Semantic ranking for automated adversarial technique annotation in security text," in *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, 2024, pp. 49–62.

[11] H. Kim and H. Kim, "Comparative experiment on ttp classification with class imbalance using oversampling from cti dataset," *Security and Communication Networks*, vol. 2022, no. 1, p. 5021125, 2022.

[12] N. Rani, B. Saha, V. Maurya, and S. K. Shukla, "Ttpxhunter: Actionable threat intelligence extraction as ttps from finished cyber threat reports," *Digital Threats: Research and Practice*, vol. 5, no. 4, pp. 1–19, 2024.

[13] Z. Hao, C. Li, X. Fu, B. Luo, and X. Du, "Leveraging hierarchies: Hmcat for efficiently mapping cti to attack techniques," in *European Symposium on Research in Computer Security*. Springer, 2024, pp. 65–85.

[14] W. You and Y. Park, "Cyber-attack technique classification using two-stage trained large language models," *arXiv preprint arXiv:2411.18755*, 2024.

[15] V. Legoy, M. Caselli, C. Seifert, and A. Peter, "Automated retrieval of att&ck tactics and techniques for cyber threat reports," *arXiv preprint arXiv:2004.14322*, 2020.

[16] L. Lange, M. Müller, G. H. Torbati, D. Milchevski, P. Grau, S. Pujari, and A. Friedrich, "Annoctr: A dataset for detecting and linking entities, tactics, and techniques in cyber threat reports," *arXiv preprint arXiv:2404.07765*, 2024.

[17] J. Ross and J. Lasky, "Our tram large language model automates ttp identification in cti reports," https://medium.com/mitre-engenuity/our-tram-large-language-model-automates-ttp-identification-in-cti-reports-5bc0a50d4567, 2023, accessed: 2025-09-25.

[18] CyberMonitor, "Apt & cybercriminal campaign collection."

[19] ESET Research, "About eset research — welivesecurity," https://www.welivesecurity.com/en/about-eset-research/, 2025, accessed: 2025-09-25.

[20] K. Bandla, "Aptnotes," https://github.com/kbandla/aptnotes/, 2025, gitHub repository, accessed 2025-09-25.

[21] "Malpedia," https://malpedia.caad.fkie.fraunhofer.de/stats/general, 2025, accessed: 2025-09-25.

[22] D. Chicco, "Siamese neural networks: An overview," *Artificial neural networks*, pp. 73–94, 2021.

[23] H.-N. Tran, A. Aizawa, and A. Takasu, "Revisiting bi-encoder neural search: An encoding–searching separation perspective," *arXiv preprint arXiv:2408.01094*, 2024.

[24] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo, "Knowledge graph embedding for link prediction: A comparative analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 2, pp. 1–49, 2021.

[25] F. Minna and F. Massacci, "Sok: run-time security for cloud microservices. are we there yet?." *Computers & Security*, p. 103119, 2023.

[26] Z. Li, J. Zeng, Y. Chen, and Z. Liang, "Attackg: Constructing technique knowledge graph from cyber threat intelligence reports," in *European Symposium on Research in Computer Security*. Springer, 2022, pp. 589–609.

[27] K. Satvat, R. Gjomemo, and V. Venkatakrishnan, "Extractor: Extracting attack behavior from threat reports," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 598–615.

[28] G. Husari, X. Niu, B. Chu, and E. Al-Shaer, "Using entropy and mutual information to extract threat actions from cyber threat intelligence," in *2018 IEEE international conference on intelligence and security informatics (ISI)*. IEEE, 2018, pp. 1–6.

[29] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8749–8757.

[30] K. Ahmed, S. K. Khurshid, and S. Hina, "Cyberentrel: Joint extraction of cyber entities and relations using deep learning," *Computers & Security*, vol. 136, p. 103579, 2024.

[31] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem, and A. Tahir, "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources," in *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2018, pp. 129–134.

[32] P. Gao, X. Liu, E. Choi, S. Ma, X. Yang, Z. Ji, Z. Zhang, and D. Song, "Threatkg: A threat knowledge graph for automated open-source cyber threat intelligence gathering and management," *arXiv preprint arXiv:2212.10388*, 2022.

[33] J. Zhao, Q. Yan, J. Li, M. Shao, Z. He, and B. Li, "Timiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data," *Computers & Security*, vol. 95, p. 101867, 2020.

[34] G. Ayoade, S. Chandra, L. Khan, K. Hamlen, and B. Thuraisingham, "Automated threat report classification over multi-source data," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2018, pp. 236–245.

[35] W. Ge and J. Wang, "Seqmask: Behavior extraction over cyber threat intelligence via multi-instance learning," *The Computer Journal*, vol. 67, no. 1.

[36] W. Ge, J. Wang, T. Lin, B. Tang, and X. Li, "Explainable cyber threat behavior identification based on self-adversarial topic generation," *Computers & security*, vol. 132, p. 103369, 2023.

[37] C. for Threat-Informed Defense, "TRAM: Threat Intelligence and Response Models," Accessed: 2024-01-22. [Online]. Available: https://github.com/center-for-threat-informed-defense/tram

[38] C. Liu, J. Wang, and X. Chen, "Threat intelligence att&ck extraction based on the attention transformer hierarchical recurrent neural network," *Applied Soft Computing*, vol. 122, p. 108826, 2022.

[39] Y. You, J. Jiang, Z. Jiang, P. Yang, B. Liu, H. Feng, X. Wang, and N. Li, "Tim: threat context-enhanced ttp intelligence mining on unstructured threat data," *Cybersecurity*, vol. 5, no. 1, p. 3, 2022.

[40] P. M. Alves, P. Geraldo Filho, and V. P. Gonçalves, "Leveraging bert's power to classify ttp from unstructured text," in *2022 Workshop on Communication Networks and Power Systems (WCNPS)*. IEEE, 2022, pp. 1–7.

[41] J. Yan, Z. Du, J. Li, S. Yang, J. Li, and J. Li, "A threat intelligence analysis method based on feature weighting and bert-bigru for industrial internet of things," *Security and Communication Networks*, vol. 2022, no. 1, p. 7729456, 2022.

[42] L. Li, C. Huang, and J. Chen, "Automated discovery and mapping att&ck tactics and techniques for unstructured cyber threat intelligence," *Computers & Security*, vol. 140, p. 103815, 2024.

[43] Y.-T. Huang, R. Vaitheeshwari, M.-C. Chen, Y.-D. Lin, R.-H. Hwang, P.-C. Lin, Y.-C. Lai, E. H.-K. Wu, C.-H. Chen, Z.-J. Liao *et al.*, "Mitretrieval: Retrieving mitre techniques from unstructured threat reports by fusion of deep learning and ontology," *IEEE Transactions on Network and Service Management*, vol. 21, no. 4, pp. 4871–4887, 2024.

[44] R. Fayyazi, R. Taghdimi, and S. J. Yang, "Advancing ttp analysis: Harnessing the power of large language models with retrieval augmented generation," in *2024 Annual Computer Security Applications Conference Workshops (ACSAC Workshops)*. IEEE, 2024, pp. 255–261.

[45] M. T. Alam, D. Bhusal, Y. Park, and N. Rastogi, "Looking beyond iocs: Automatically extracting attack patterns from external cti," in *Proceedings of the 26th international symposium on research in attacks, intrusions and defenses*, 2023, pp. 92–108.

[46] B. Abdeen, E. Al-Shaer, A. Singhal, L. Khan, and K. Hamlen, "Smet: Semantic mapping of cve to att&ck and its application to cybersecurity," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2023, pp. 243–260.
[47] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco, "Time for action: Automated analysis of cyber threat intelligence in the wild," *arXiv preprint arXiv:2307.10214*, 2023.
[48] M. Chen, K. Zhu, B. Lu, D. Li, Q. Yuan, and Y. Zhu, "Aecr: Automatic attack technique intelligence extraction based on fine-tuned large language model," *Computers & Security*, vol. 150, p. 104213, 2025.
[49] Y. Fengrui and Y. Du, "Few-shot learning of ttps classification using large language models," 2024.
[50] M. Xu, H. Wang, J. Liu, Y. Lin, C. X. Y. Liu, H. W. Lim, and J. S. Dong, "Intelex: A llm-driven attack-level threat intelligence extraction framework," *arXiv preprint arXiv:2412.10872*, 2024.
[51] Y. Zhang, T. Du, Y. Ma, X. Wang, Y. Xie, G. Yang, Y. Lu, and E.-C. Chang, "Attackg+: Boosting attack knowledge graph construction with large language models," *arXiv preprint arXiv:2405.04753*, 2024.
[52] R. Fieblinger, M. T. Alam, and N. Rastogi, "Actionable cyber threat intelligence using knowledge graphs and large language models," in *2024 IEEE European symposium on security and privacy workshops (EuroS&PW)*. IEEE, 2024, pp. 100–111.
[53] H. Cuong Nguyen, S. Tariq, M. Baruwal Chhetri, and B. Quoc Vo, "Towards effective identification of attack techniques in cyber threat intelligence reports using large language models," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 942–946.
[54] D. Hamzic, F. Skopik, M. Landauer, M. Wurzenberger, and A. Rauber, "Ttp classification with minimal labeled data: A retrieval-based few-shot learning approach," in *International Conference on Availability, Reliability and Security*. Springer, 2025, pp. 387–408.
[55] A. Joy, M. Chandane, Y. Nagare, and F. Kazi, "Threat intelligence extraction framework (tief) for ttp extraction," *Journal of Cybersecurity and Privacy*, vol. 5, no. 3, p. 63, 2025.
[56] Y. Schwartz, L. Ben-Shimol, D. Mimran, Y. Elovici, and A. Shabtai, "Llmcloudhunter: Harnessing llms for automated extraction of detection rules from cloud-based cti," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1922–1941.
[57] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang, "Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model," *Computers & Security*, vol. 145, p. 103999, 2024.

XI. Appendix

A. Cross-Source Redundancy and Complementarity

To assess whether combining multiple CTI report sources primarily adds new evidence or simply duplicates existing reports, we measured cross-source overlap using SHA-256 fingerprints over sanitized report text. We find limited overlap across sources (only a small number of duplicate documents across the full collection), with the largest duplicate groups concentrated in a few source pairs (notably MITRE reports–Malpedia and APT notes–APT & Cybercriminals). Overall, the combined CTI document corpus is largely complementary rather than a simple re-hosting of identical reports.

B. TTP-SiamAlign architecture

The TTP-SiamAlign pipeline is composed of 4 steps: (A) A labeled sentence–TTP dataset is used as supervised training data. (B) A siamese bi-encoder with shared weights is trained to project CTI sentences and ATT&CK technique descriptions into a shared embedding space using contrastive learning. (C) The trained model is applied to unseen CTI texts, where sentences are matched against candidate technique descriptions by cosine similarity. (D) High-confidence matches are exported as pseudo-labeled sentence–TTP pairs, which can be used to expand the training set.

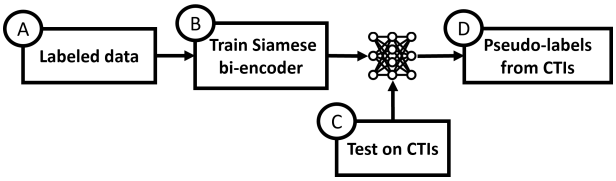

Fig. 3. Overview of the siamese bi-encoder (*TTP-SiamAlign*) pipeline for sentence–TTP matching.

C. Export and Allocation

Recovered sentence–TTP matches from the CTI corpus can be exported in multiple ways before being used for augmentation. Because a single technique may be matched by many candidate sentences of varying confidence and redundancy, the export policy directly affects the precision, size, and diversity of the resulting augmentation pool. We therefore evaluate two export strategies.

a) Global-best-per-TTP.: In this setting, we retain only a single sentence for each ATT&CK technique, namely the sentence with the highest similarity score across the entire CTI corpus. This yields a compact augmentation set that emphasizes precision and minimizes redundancy. Because only the strongest match per label is exported, this strategy is conservative and particularly suitable when the goal is to reduce the risk of noisy supervision.

b) Best-per-TTP-per-document.: In this setting, we retain the highest-scoring sentence for each technique within each source document. Consequently, the same technique may contribute multiple exported sentences, provided that they originate from different reports. Compared with the global-best setting, this policy produces a substantially larger candidate pool and preserves greater linguistic and contextual diversity, as different reports often describe the same behavior using different wording, detail, or narrative framing.

These two export strategies reflect a deliberate trade-off between precision and diversity. *Global-best-per-TTP* favors a small set of highly confident exemplars, whereas *best-per-TTP-per-document* exposes the model to broader variation in how techniques are expressed in practice. Evaluating both allows us to isolate whether augmentation gains are driven primarily by candidate

TABLE IX
OVERVIEW OF DATASETS FOR TTP CLASSIFICATION.

Paper	Symantec	Debrin	Apt notes	APT&Cyber	WeliveSecurity	rcATT	ATT&CK reports	ATT&CK procedures	ATT&CK descr	CAPEC	TRAM	AnnoCTR	*Proprietary*	*Reproducible*	*Other*
AttacKG [26]	○	○	○	○	○	○	●	●	●	○	○	○	○	●	○
Extractor [27]	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○
TTPDrill [3]	●	○	○	○	○	○	○	○	○	○	○	○	○	●	○
ActionMiner [28]	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
CASIE [29]	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
CyberEntRel [30]	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
EX-Action [5]	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○
[31]	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
ThreatKG [32]	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
TIMiner [33]	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
rcATT [15]	○	○	○	○	○	●	○	○	○	○	○	○	○	●	○
[34]	○	○	○	○	○	○	●	○	○	○	○	○	○	○	●
SeqMask [35]	●	●	○	○	○	●	○	●	○	○	○	○	○	●	●
[36]	●	●	○	○	○	●	○	●	○	○	○	○	○	●	●
TRAM [37]	○	○	○	○	○	○	○	○	○	○	●	○	○	●	○
ATHRNN [38]	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
TTPHunter [6]	○	○	○	○	○	●	●	●	●	○	○	○	○	●	●
TIM [39]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
[40]	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○
[41]	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○
[11]	○	○	○	○	○	○	○	●	○	○	●	○	○	○	○
TTPXHunter [12]	○	○	○	○	○	○	○	●	○	○	○	○	○	○	●
[14]	○	○	○	○	○	○	○	○	●	○	●	○	○	○	○
Alert [7]	○	○	○	○	○	○	○	○	○	○	●	○	○	●	○
[42]	○	○	○	○	○	○	○	●	○	○	●	○	○	●	○
CTI-to-MITRE [8]	○	○	○	○	○	○	○	●	●	●	●	○	○	●	○
HMCAT [13]	○	○	○	○	○	○	○	●	●	●	○	○	○	●	○
MITREtriva [43]	○	○	○	●	○	●	○	○	○	○	○	○	○	●	●
[44]	○	○	○	○	○	○	○	●	●	○	○	○	○	●	○
LADDER [45]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
Kumaras [10]	○	○	○	○	●	○	●	○	○	○	●	○	○	●	●
Smet [46]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
aCTIon [47]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
AECR [48]	○	○	○	●	●	○	○	●	●	●	○	○	○	●	●
[49]	○	○	○	○	○	○	○	●	●	○	○	○	○	●	○
IntelEX [50]	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●
Attackg+ [51]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
[52]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
AnnoCTR [16]	○	○	○	○	○	○	○	○	○	○	○	●	○	●	○
SOK [2]	○	○	○	○	○	○	○	○	○	○	●	●	○	●	○
[53]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
TTPFShot [54]	○	○	○	○	○	●	●	●	●	○	○	○	○	●	●
TIEF [55]	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○
CTI bench [4]	○	○	○	○	○	○	○	●	○	○	○	○	○	●	○
LLMCloudHunter [56]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●
LLM-TIKG [57]	●	○	○	○	○	○	○	○	○	○	○	○	○	●	●
CTI-HAL [9]	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●

Each row corresponds to a published paper, and each column to a dataset source. A full circle (●) indicates that the dataset was used, an empty circle (○) that it was not, and "Upon request" denotes restricted availability. The last two columns summarize whether the dataset is reproducible and whether additional data sources were employed. Note that the inclusion of the data sources in the data sources column depends on the fact that the specific source must be used by at least two papers.

quality or by the increased coverage and variability obtained from a larger exported pool.

D. Allocation Strategies

After export, the resulting candidate pool must still be allocated to the benchmark training set. This is necessary because label frequencies in both the original benchmark data and the exported augmentation pool are often highly imbalanced. Without allocation control, augmentation may disproportionately benefit already frequent labels while contributing little to underrepresented ones. We therefore study four allocation strategies that differ in how aggressively they preserve or correct the original class skew.

a) Global.: The *global* strategy adds all exported sentence–TTP pairs without imposing any label-wise budget. This maximizes augmentation volume and preserves the natural distribution of the exported pool. However, because frequent techniques typically produce more candidates, this strategy can amplify head-label dominance and may further skew the training distribution. Figure 4 illustrates this behavior.

b) Uniform.: The *consistent* strategy allocates a fixed and uniform number of augmented pairs per label, subject to candidate availability. In contrast to *till-max*, which depends on the observed class maximum, this strategy enforces the same augmentation target across labels and therefore provides more direct control over the number of added examples per class. Its purpose is to reduce imbalance while maintaining a predictable and comparable augmentation budget across experiments. An example is shown in Figure 6.

c) Cap-by-max.: The *till-max* strategy allocates augmentation preferentially to underrepresented labels until they reach the size of the largest class in the original training split. Labels that already match or exceed this maximum receive no additional examples. This strategy acts as a balancing mechanism: it allows minority classes to grow while preventing uncontrolled expansion of already frequent labels. Figure 5 shows the resulting effect on the label distribution.

d) Performance-driven.: The *performance-driven* strategy uses validation performance to guide allocation. Rather than applying a single fixed rule to all labels, it prioritizes augmentation settings that empirically improve downstream classification. In practice, this strategy treats allocation as a data-centric hyperparameter: additional examples are directed toward configurations that appear most beneficial under validation, instead of being determined solely by class counts. This makes it the most adaptive strategy, but also the most dependent on reliable validation feedback. Figure 7 provides an illustration.

Taken together, these four strategies span a spectrum from unconstrained augmentation to strongly controlled, label-aware allocation. The *global* setting prioritizes supervision volume, *till-max* emphasizes class balancing relative to the observed benchmark skew, *consistent* enforces a uniform augmentation budget across labels, and *performance-driven* allocates supervision where it appears most useful empirically. Comparing them allows us to distinguish whether augmentation gains arise simply from adding more data or from adding data in a distributionally controlled manner.

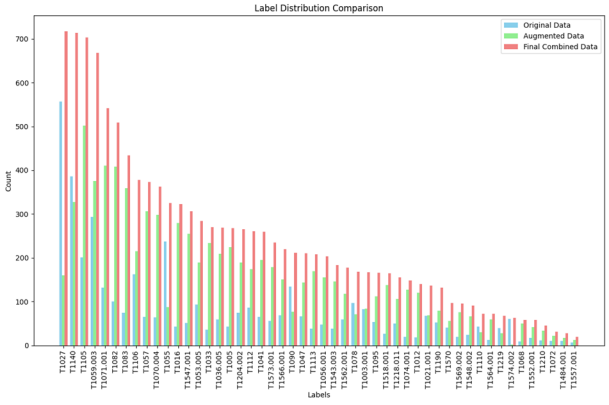

Fig. 4. Illustration of the *global* allocation strategy, where all exported augmentation pairs are added without label-wise constraints.

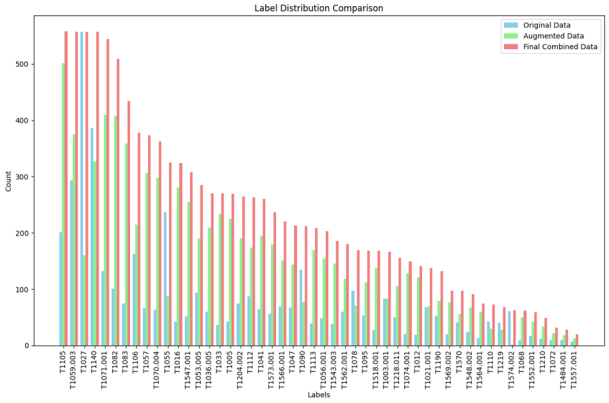

Fig. 5. Illustration of the *till-max* allocation strategy, where underrepresented labels are augmented until they reach the size of the largest original class.

E. Preprocessing

CTI reports frequently contain Indicators of Compromise (IoCs), such such as cryptographic hashes, IP addresses, URLs, registry keys, and file paths—which can artificially inflate lexical variability without contributing to the semantic identification of ATT&CK techniques. Following prior work on CTI-to-ATT&CK

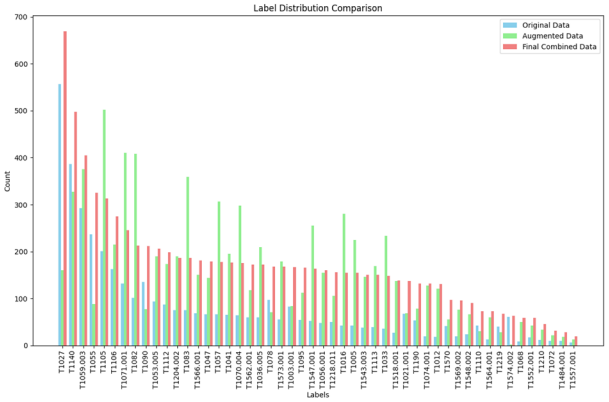

Fig. 6. Illustration of the *consistent* allocation strategy, where each label receives the same fixed augmentation budget, subject to candidate availability.

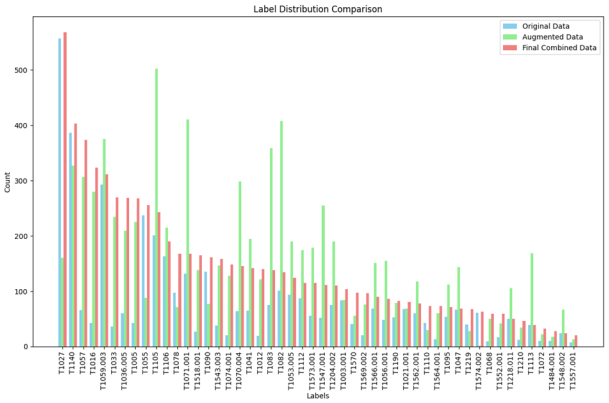

Fig. 7. Illustration of the *performance-driven* allocation strategy, where augmentation is guided by validation performance rather than by a purely count-based rule.

classification [8], [13], we evaluate a pattern-based normalization pipeline that replaces matched IoCs with typed placeholders (Table XI). The regex-based pipeline ensures consistent detection across indicator formats (e.g., `127.0.0.1` → `<IPv4>`).

Table X reports the effect of IoC normalization on downstream classification performance. IoC normalization does not yield consistent benefits. On TRAM, its impact is negligible across all metrics, indicating that preprocessing neither meaningfully improves nor degrades performance in that setting. On AnnoCTR, however, normalization leads to a systematic decline, suggesting that replacing raw indicators with placeholders can remove lexical or contextual signals that remain informative for fine-grained technique discrimination, especially in smaller datasets.

Overall, these results indicate that IoC normalization should not be assumed to be beneficial by default. Instead, its utility is dataset-dependent and should be

validated empirically under controlled experimental conditions before being adopted.

TABLE X
IMPACT OF PREPROCESSING ON THE BEST-PERFORMING MODEL ACROSS TRAM AND ANNOCTR. WE REPORT THE PERCENTAGE DIFFERENCE RELATIVE TO THE BEST-PERFORMING MODEL.

Setting	**TRAM** [17]			**AnnoCTR** [16]		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Baseline	77.20	62.80	66.57	70.49	51.03	56.62
Preprocesses	77.10	61.90	66.11	66.51	45.61	51.21
Improvement	-0.13%	-1.43%	-0.69%	-5.65%	-10.63%	-9.56%

F. Hyperparameters for sentence-ttp classification

G. Global/local CTI-extracted sentence

Table XIII indicates that, under *global-best-per-TTP*, the most reliable gains are obtained with conservative allocation. In particular, *uniform* allocation achieves the best overall trade-off despite adding only a small amount of data (TRAM: +2.12%, AnnoCTR: +17.83%). In this setting, precision remains comparable to the baseline while recall improves, yielding consistent F1 gains on both datasets.

In contrast, Table XIV shows that *best-per-TTP-per-document* can inflate the pseudo-labeled pool by orders of magnitude (e.g., TRAM: +147.84%; AnnoCTR: +541.38%), but the added diversity substantially increases noise. Naively exporting large volumes of pseudo-labels tends to degrade precision and does not translate into proportional improvements in downstream performance, illustrating that CTI mining is not a "more is better" regime.

Across both export regimes, these results emphasize that the effectiveness of CTI-derived augmentation depends critically on allocation controls. Strict per-label budgeting (uniform or capped/performance-driven policies) prevents head-label dominance and limits pseudo-label noise, enabling CTI-derived evidence to complement supervised training rather than overwhelm it.

H. Examples of different augmentation strategies.

I. Full comparison across classification models

TABLE XI
REGEX- AND GAZETTEER-BASED PREPROCESSING RULES APPLIED TO CTI SENTENCES.

Rule name	**Description**	**Example match**
URL	Web URLs (including HTTP/HTTPS links)	`http://ex.com`
Domain	Fully qualified domain names	`example.com`
Email	Email addresses	`mail@es.com`
IPv4 Address	IPv4 network addresses	`192.168.1.1`
IPv6 Address	IPv6 network addresses	`2001:0db8:...`
File Path	Windows or Unix-style file system paths	`C:\Us\bob\e.exe`
Filename	Executable or document filenames	`payload.dll`, `script.ps1`
Registry Key	Windows registry keys	`HKCU\Windows\Run`
File Hash	Cryptographic file hashes (MD5, SHA-1, SHA-256, etc.)	`66eff......76db`
CVE Identifier	Common Vulnerabilities and Exposures identifiers	`CVE-2017-11882`
MAC Address	Network interface MAC addresses	`00:1A:2B:3C:4D:5E`
Named Pipe	Windows named pipes used for IPC	`\\.\pipe\svcctl`
Admin Share	Windows administrative network shares	`\\HOST\C$`
Encryption / Encoding Algorithm	Cryptographic or encoding algorithm names (gazetteer-based)	`Base64`, `AES`, `XOR`
Communication Protocol	Network communication protocols (gazetteer-based)	`HTTP`, `SMTP`, `DNS`
Data Object	Sensitive data objects (gazetteer-based)	`clipboard`, `password`, `keystrokes`

TABLE XII
TRAINING HYPERPARAMETERS FOR CLASSIFICATION-BASED AND GENERATIVE TTP EXTRACTION MODELS.

Category	**Classification (Multi-label Fine-Tuning)**	**Generation (LLM-based)**
Base Model	Pre-trained encoder (BERT/RoBERTa-family, CTI-specific variants)	Llama3.1-Instruct-8B (instruction-tuned LLM)
Task Formulation	Multi-label sentence classification (one logit per label)	Instruction-based generation of TTP IDs (parsed from output text)
Optimizer	AdamW	AdamW (LoRA fine-tuning)
Learning Rate	2e-05	1×10^{-5} (embedding layers), 2×10^{-5} (other layers)
Loss Function	BCEWithLogitsLoss (optional positive class weighting)	Cross-entropy (causal LM objective)
Batch Size	16	4
Epochs	max 100 (early stopping enabled)	3
Gradient Clipping	Max norm = `MAX_GRAD_NORM`	Applied (LoRA training)
Class Imbalance Handling	Optional `pos_weight` in BCE loss	None (implicitly learned during fine-tuning)
Early Stopping	Based on validation loss (patience = `15`)	Not used (fixed epochs)
Parameter Update Scope	Full model	LoRA adapters only
Precision	FP32	16-bit weights (no quantization)

TABLE XIII
EFFECT OF AUGMENTATION STRATEGIES ON TRAM AND ANNOCTR USING CTI-DERIVED SENTENCES SELECTED WITH THE GLOBAL-BEST-PER-TTP RULE.

Aug.	**TRAM**				**AnnoCTR**			
	Pairs	**P**	**R**	**F1**	**Pairs**	**P**	**R**	**F1**
(0) Base	12017	77.20	62.80	66.57	2934	70.49	51.03	56.62
(1) Global	2315	77.67	59.65	64.71	2289	74.83	51.00	57.89
(2) Long-tail	2310	79.04	58.24	64.88	2284	77.19	49.69	57.39
(3) Uniform	255	76.84	64.38	67.63	523	71.09	55.72	59.92
(4) Perf.-driven	255	77.92	63.21	67.30	308	72.40	53.15	58.70

TABLE XIV
CONTRIBUTION OF AUGMENTATION STRATEGIES ON TRAM AND ANNOCTR WITH CTI-DERIVED SENTENCES (BEST-PER-TTP-PER-DOCUMENT).

Aug.	**TRAM**				**AnnoCTR**			
	Pairs	**P**	**R**	**F1**	**Pairs**	**P**	**R**	**F1**
(0) Base	12017	77.20	62.80	66.57	2934	70.49	51.03	56.62
(1) Global	17789	73.31	50.38	57.57	15884	70.51	48.25	55.22
(2) Long-tail	16638	76.97	51.13	59.43	14043	–	–	–
(3) Uniform	5127	70.18	68.85	67.64	4295	62.22	58.65	58.36
(4) Perf.-driven	4560	71.84	67.91	67.61	586	69.05	52.81	57.39

Method	**Original**	**Augmented**
EDA+BT [11]	"... a web **shell** on a vulnerable Microsoft Exchange..."	"... a web **carapace** on a vulnerable Microsoft Exchange..."
TTP X HUNTER [12]	"For the second account, which also had Global Administrator permissions, the threat actors leveraged RDP for access **into** the account."	"For the second account, which also had Global Administrator permissions, the threat actors leveraged RDP for access **through the** account."
HMCAT [13]	'Malicious files are renamed as pirated software or gaming software to trick gamers.'	'Phishing attachments employ double extensions like invoice.pdf.exe to conceal executable content from end users.'

TABLE XV
CONTRIBUTION OF AUGMENTATION STRATEGIES ON TRAM AND ANNOCTR.

Models	**TRAM** [17]			**AnnoCTR** [16]		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
CyBERT	74.54 ±2.91	45.99 ±7.91	54.01 ±6.24	75.13 ±3.42	41.06 ±3.52	49.90 ±3.23
CySecBERT	80.50 ±3.39	51.79 ±10.01	60.08 ±7.16	73.02 ±3.26	48.73 ±3.15	55.66 ±2.57
DarkBERT	79.02 ±3.60	58.06 ±4.65	64.34 ±2.99	73.47 ±3.32	45.86 ±2.89	53.43 ±2.61
SecBERT	76.07 ±1.20	51.75 ±1.62	60.06 ±1.49	**80.01** ±3.13	41.96 ±3.18	51.23 ±2.99
SecRoBERTa	**81.97** ±0.88	44.13 ±3.15	54.70 ±2.72	78.06 ±1.67	34.05 ±1.78	44.51 ±1.44
SecureBERT	77.20 ±1.73	**62.80** ±5.23	**66.57** ±3.81	75.38 ±4.10	45.13 ±4.62	53.23 ±3.18
BERT_(Base, Cased)	78.82 ±1.89	53.98 ±6.42	61.68 ±4.74	71.18 ±3.50	38.09 ±2.76	46.37 ±2.06
BERT_(Base, Uncased)	77.72 ±2.86	47.55 ±8.46	56.61 ±6.44	74.36 ±2.29	41.18 ±3.99	49.48 ±3.71
RoBERTa$_{(Base)}$	77.49 ±3.08	55.10 ±7.96	61.86 ±5.64	71.01 ±2.97	41.41 ±4.03	49.32 ±3.49
RoBERTa$_{(Large)}$	79.89 ±3.05	56.70 ±4.63	63.97 ±3.21	70.49 ±4.57	**51.03** ±2.88	**56.62** ±2.91
XLM-RoBERTaBase	77.71 ±2.42	53.23±7.93	60.42 ±4.93	71.62 ±4.47	39.07 ±4.30	47.05 ±3.83
SciBERTCased	78.27 ±3.27	55.27 ±5.64	62.58 ±3.63	72.93 ±5.06	44.29 ±2.19	52.06 ±2.40
SciBERTUncased	78.10 ±4.38	55.57 ±9.57	62.09 ±6.07	73.79 ±2.94	44.20 ±3.75	52.07 ±2.87

TABLE XVI
CONTRIBUTION OF AUGMENTATION STRATEGIES ON TRAM AND ANNOCTR. "ADDED PAIRS" REFERS TO NEWLY GENERATED SENTENCE–TTP TRAINING PAIRS ADDED TO THE ORIGINAL TRAINING SPLIT.

Augmentation strategy	**TRAM** [17]				**AnnoCTR** [16]			
	Train pairs	**Prec.**	**Rec.**	**F1**	**Train pairs**	**Prec.**	**Rec.**	**F1**
(0) Baseline	12017	77.20±1.73	62.80±5.23	66.57±3.8	2934	70.49±4.57	51.03±2.88	56.62±2.91
(1) Global	+7929 (+65.98%)	75.26±2.41	61.33±4.79	65.63±2.44	+6535 (+222.73%)	68.89±4.03	59.88±6.08	61.58±2.90
(2) Long-tail	+7106 (+59.13%)	74.20±2.95	64.20±6.05	66.99±2.88	+5446 (+185.62%)	68.67±5.07	53.47±6.49	57.36±4.07
(3)Uniform								
(3a) 20%	+3748 (+31.19%)	75.25±2.07	65.80±4.45	68.16±2.18	+2879 (+98.13%)	66.88±2.66	55.46±4.09	58.37±2.83
(3b) 30%	+5045 (+41.98%)	73.41±2.08	66.24±8.68	67.15±4.15	+3525 (+120.14%)	69.69±3.22	55.38±5.60	59.42±4.09
(3c) 40%	+6012 (50.03%)	74.43±1.93	67.37±3.31	68.52±1.79	+3995 (136.16%)	69.42±3.51	53.92±4.18	58.28±3.16
(4) Performance driven								
(4a) 20% cap	+2454 (+20.42%)	74.20±2.48	62.11±4.27	65.96±2.32	+585 (+19.94%)	72.14±2.89	53.18±2.91	58.70±2.39
(4b) 30%cap	+3680 (+30.62%)	75.36±1.88	65.30±6.36	67.87±3.62	+879 (+29.96%)	69.05±4.44	53.41±3.69	57.79±3.09
(4c) 40%cap	+4907 (+40.83%)	72.65±1.78	68.63±4.26	68.30±2.13	+1171 (+39.91%)	68.26±1.71	54.35±2.96	58.12±2.00