Carl von Ossietzky
Universität
Oldenburg

SoK: Automated TTP Extraction from CTI Reports – Are We There Yet?

Marvin Büchel, Tommaso Paladini, Stefano Longari, Michele Carminati, Stefano Zanero, Hodaya Binyamini, Gal Engelberg, Dan Klein, Giancarlo Guizzardi, Marco Caselli, Andrea Continella, Maarten van Steen, Andreas Peter, and Thijs van Ede

POLITECNICO
MILANO 1863

UNIVERSITY OF TWENTE.

accenture SIEMENS NEC

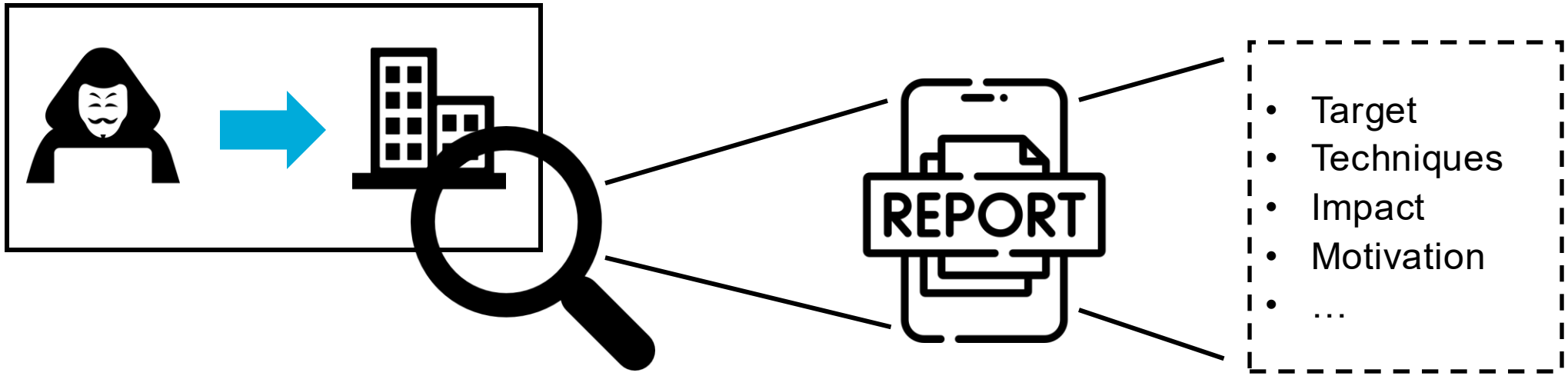
marvin.buechel@uol.de

SSI

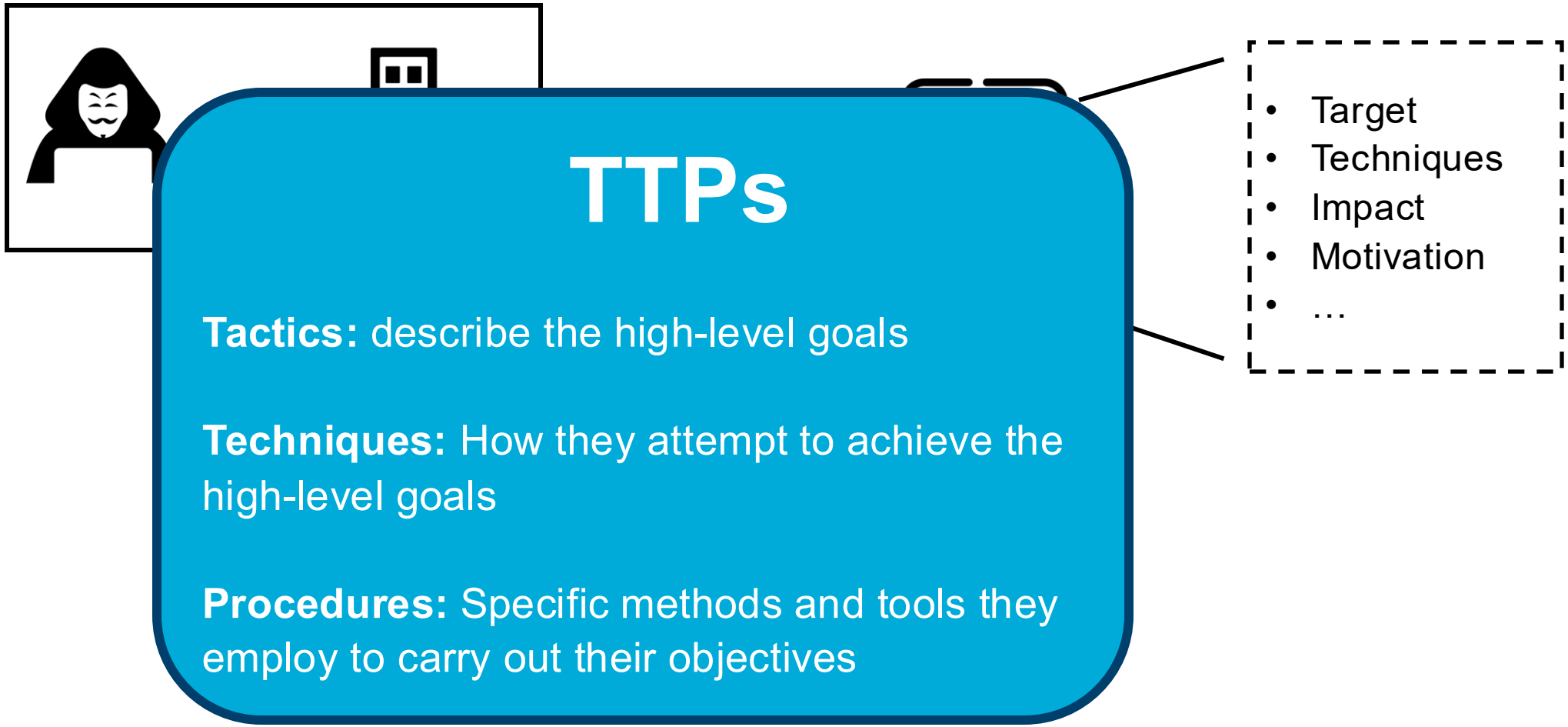
Cyber Threat Intelligence (CTI)

CTI reports containing important information about recent cyber attacks…

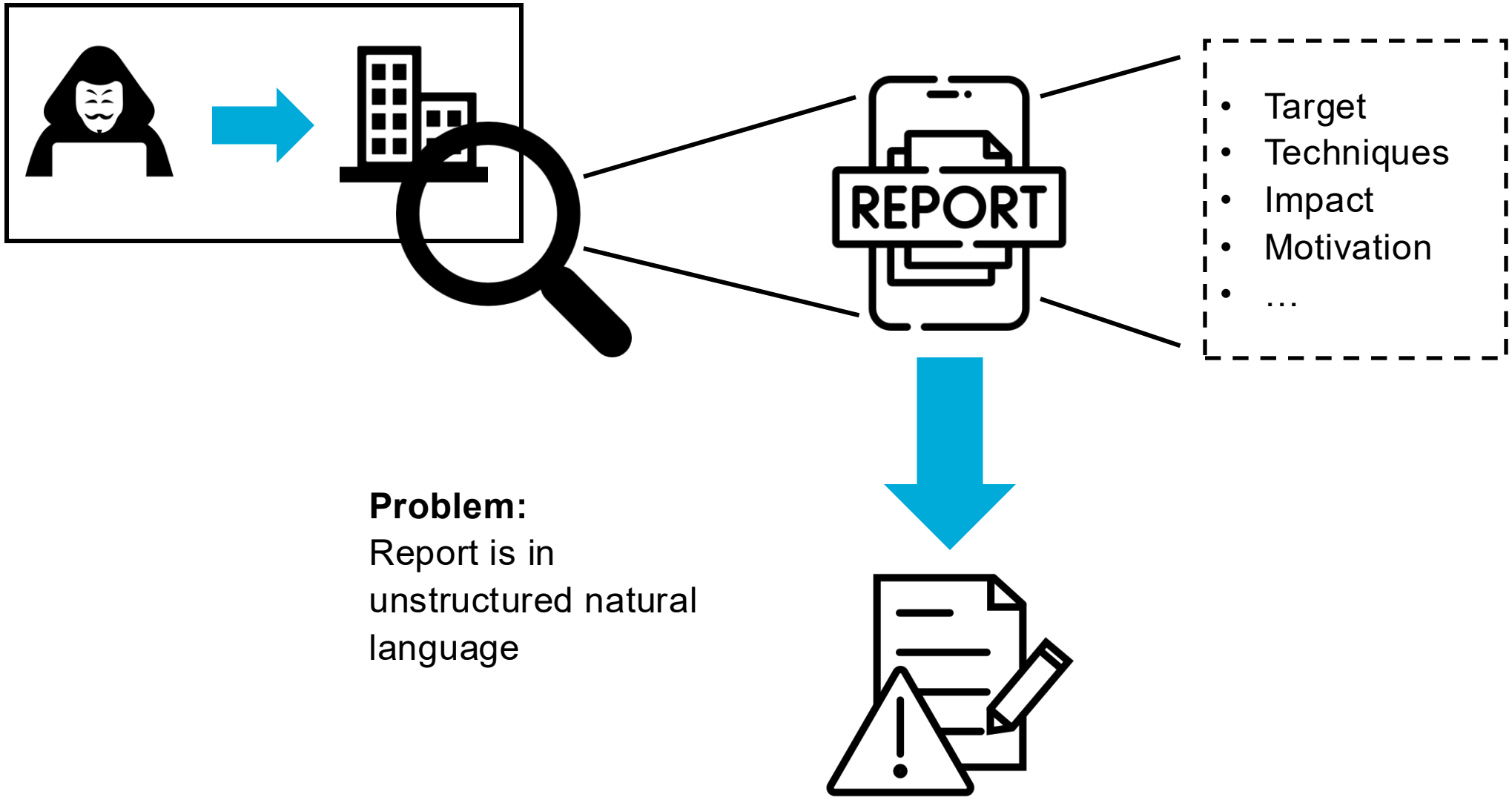
Cyber Threat Intelligence (CTI)



Cyber Threat Intelligence (CTI)

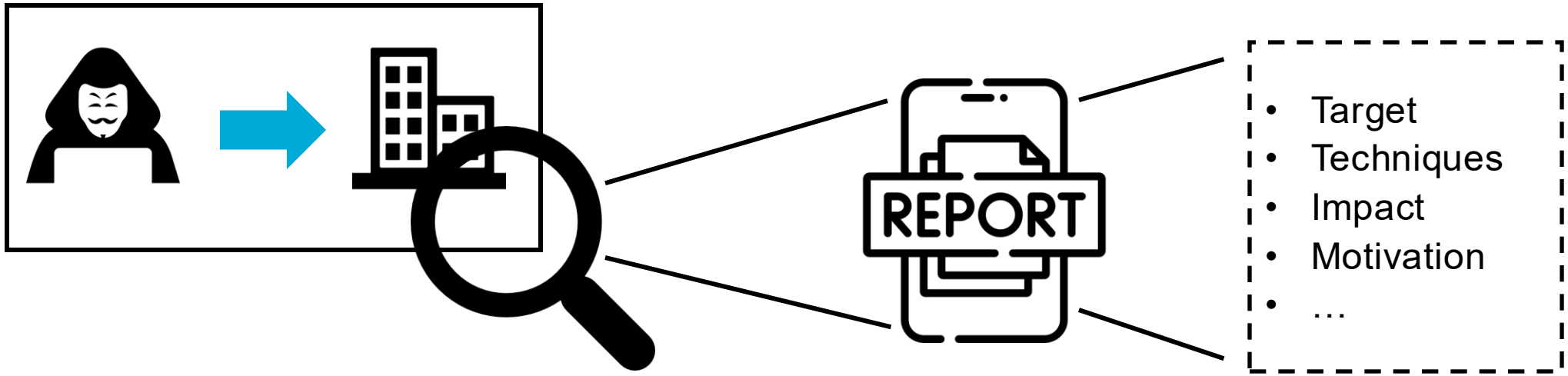


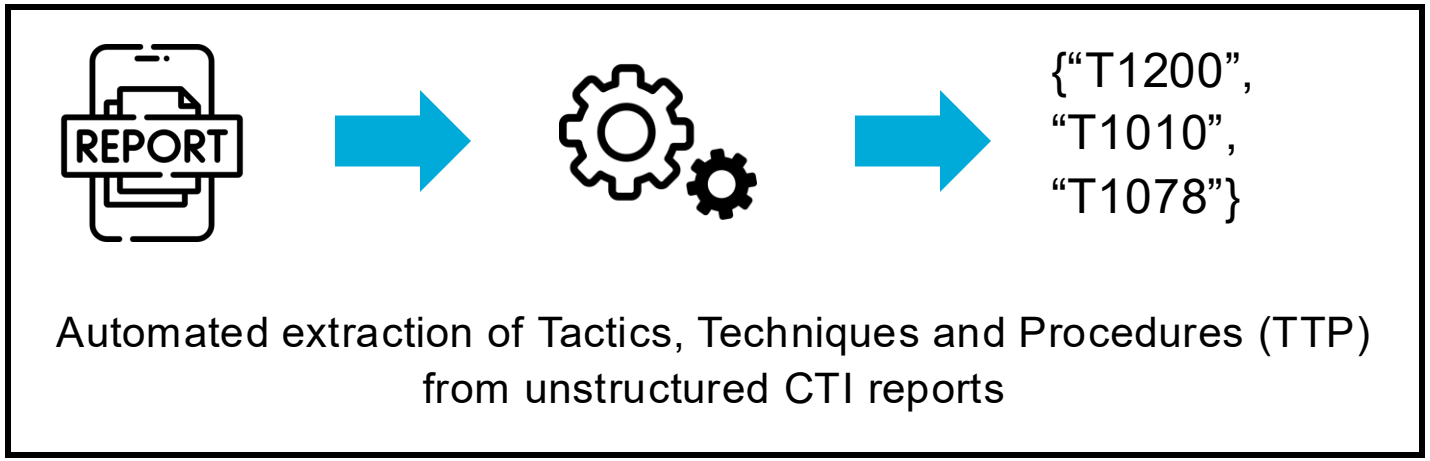
Cyber Threat Intelligence (CTI)



Problem:
Report is in unstructured natural language

Cyber Threat Intelligence (CTI)





The MITRE ATT&CK® Ontology

- De-facto standard for categorizing TTPs
- A globally accessible knowledge base of TTPs based on real-world observations
- Unique IDs for Tactics and Techniques
- Creates a common language for security professionals and enables automated processing

Goal: Map *'The attacker modified the Windows Registry' -> ATT&CK ID T1112*

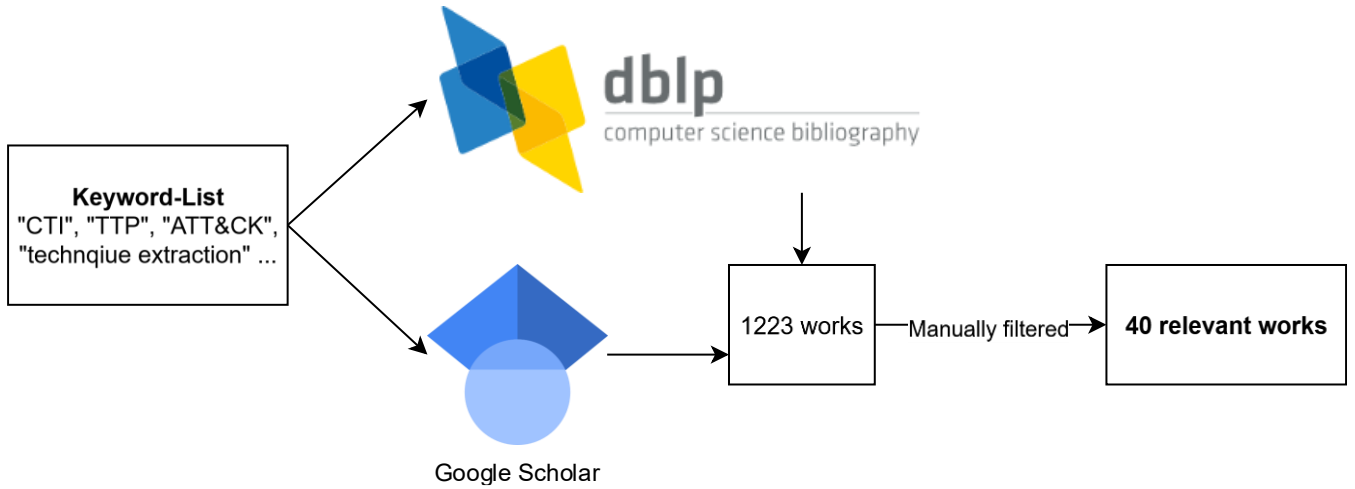
Literature Review

Objective: To systematize all existing NLP-based solutions for automatic TTP extraction.

Literature Review

Objective: To systematize all existing NLP-based solutions for automatic TTP extraction.

Process:



Our Systematization of Knowledge

- **40 relevant works published**

Named Entity Recognition (NER): Advanced rule-base "word search"

Classification: Data-driven classification approach (often BERT-like models)

Generation: Generative Large Language Model (LLM)

Our Systematization of Knowledge

- **40 relevant works published**

Named Entity Recognition (NER): Advanced rule-base "word search"

Classification: Data-driven classification approach (often BERT-like models)

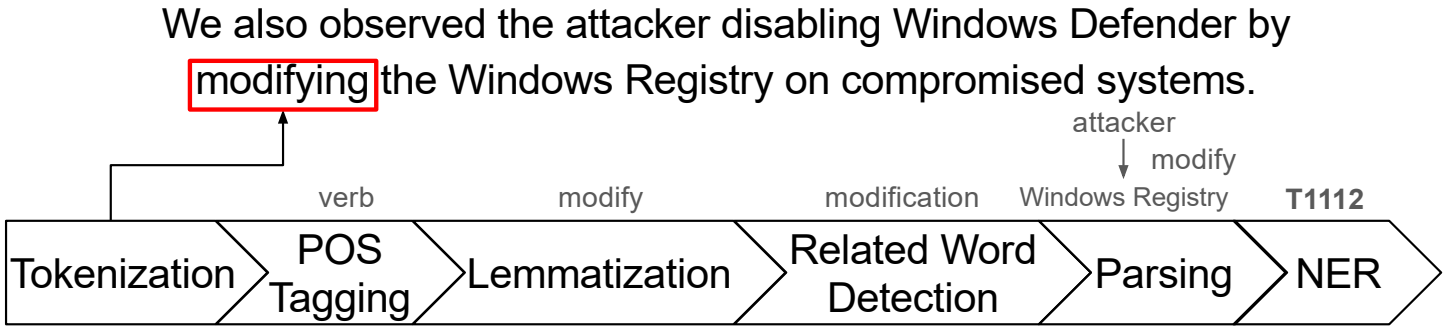
Generation: Generative Large Language Model (LLM)

- **Identified 16 fundamental Natural Language Processing (NLP) methods**

Approach Methods NER

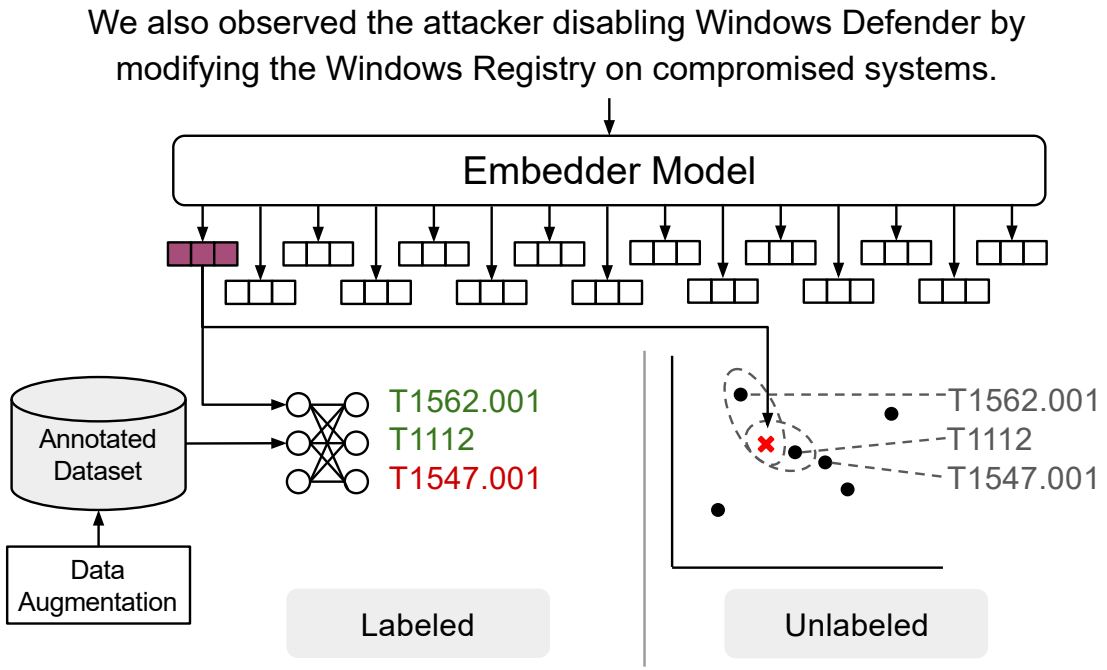
Each work consists of a composition of NLP methods within the general approach.

NER Approach Example:



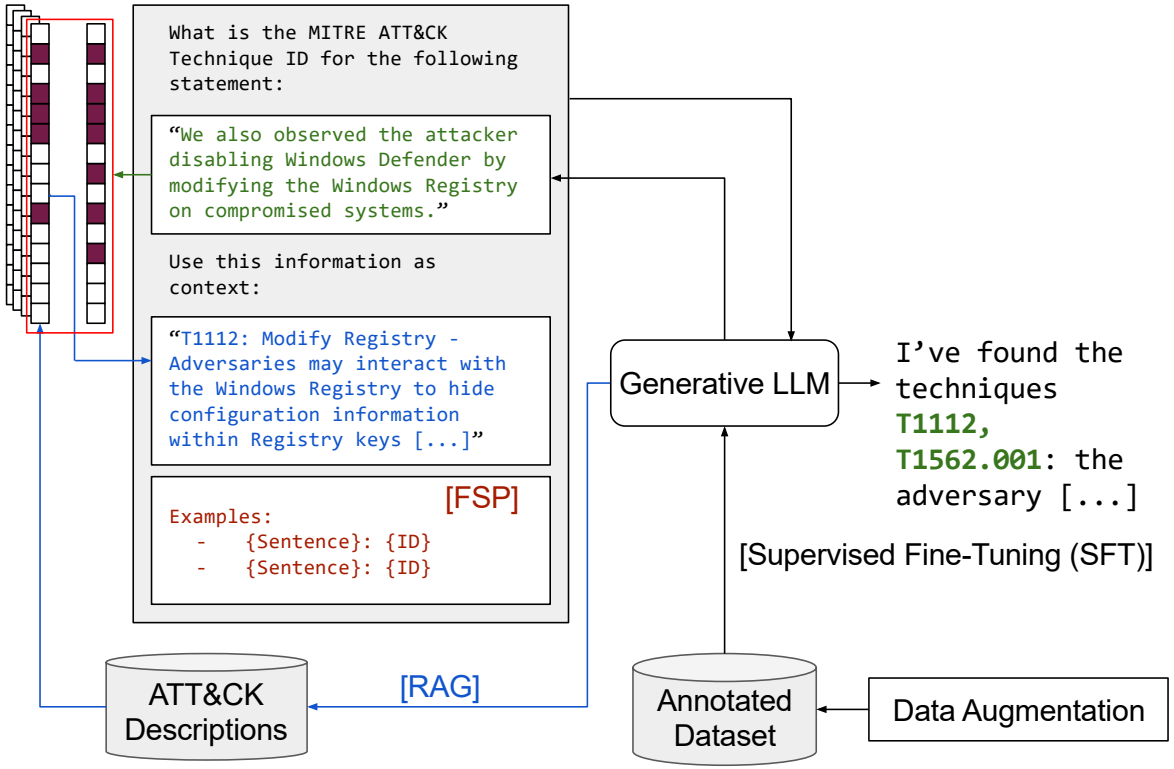
Approach Methods Classification

Structure Classification Approach:



Approach Methods Classification

Structure Generation Approach:



Literature Trends - NER

Table 1: Components implemented by related work.

Type	Approach	Tok.	POS	Lem.	Rel.	Par.	NER	Data	Ontology
Semantic	AttacKG [58]	❍	❍	❍	×	❍	●	Custom	Custom
	Extractor [85]	●	●	●	❍	●	●	[20, 66]	Custom
	TTPDrill [38]	●	●	×	●	❍	●	Custom	Custom
Hybrid	ActionMiner [39]	❍	❍	×	×	❍	●	Custom	Custom
	CASIE [86]	❍	❍	❍	×	●	●	Custom	Custom
	CyberEntRel [3]	❍	×	×	❍	×	●	Custom	Custom
	EX-Action [103]	●	❍	×	×	❍	●	Custom	ATT&CK
	Ghazi et al. [32]	❍	❍	×	×	×	❍	Custom	Custom
	ThreatKG [27]	●	×	×	×	●	●	Custom	Custom
	TIMiner [106]	❍	×	❍	●	❍	●	Custom	Custom

×	action is not present or not mentioned.	Tok.	=	Tokenization	Rel.	=	Related word detection
❍	action is present, but not domain-specific.	POS	=	Part-of-Speech	Par.	=	Parsing
●	action is present and domain-specific.	Lem.	=	Lemmatization	NER	=	Named Entity Recognition

Literature Trends - NER

Table 1: Components implemented by related work.

Type	Approach	Tok.	POS	Lem.	Rel.	Par.	NER	Data	Ontology
Semantic	AttacKG [58]	❍	❍	❍	×	❍	●	Custom	Custom
	Extractor [85]	●	●	●	❍	●	●	[20, 66]	Custom
	TTPDrill [38]	●	●	×	●	❍	●	Custom	Custom
Hybrid	ActionMiner [39]	❍	❍	×	×	❍	●	Custom	Custom
	CASIE [86]	❍	❍	❍	×	●	●	Custom	Custom
	CyberEntRel [3]	❍	×	×	❍	×	●	Custom	Custom
	EX-Action [103]	●	❍	×	×	❍	●	Custom	ATT&CK
	Ghazi et al. [32]	❍	❍	×	×	×	❍	Custom	Custom
	ThreatKG [27]	●	×	×	×	●	●	Custom	Custom
	TIMiner [106]	❍	×	❍	●	❍	●	Custom	Custom

× action is not present or not mentioned.
❍ action is present, but not domain-specific.
● action is present and domain-specific.

Tok. = Tokenization
POS = Part-of-Speech
Lem. = Lemmatization
Rel. = Related word detection
Par. = Parsing
NER = Named Entity Recognition

Literature Trends - Classification

Table 2: Classification approaches for TTP extraction.

	Paper	Model	Augment. Art.	Augment. OOD	Mult.	Ont.	CTRs	Data	Gran.
Labeled	rcATT [52]	Word2Vec [65]	×	×	●	ATT.	●	Custom	Doc.
	Ayoade et al. [7]	None	×	×	●	ATT.	●	Custom	Doc.
	SeqMask [30], Ge et al. [31]	FastText [12]	×	×	●	ATT.	●	Custom	Doc.
	TRAM [82, 99]	SciBERT [11]	×	×	●	ATT.	●	[82]	Sent.
	Liu et al. [61]	Custom	×	×	●	ATT.	●	Custom	Doc.
	TTPHunter [79]	BERT-based [22]	×	×	×	ATT.	●	Custom	Sent.
	TIM [101]	SBERT [81]	×	×	×	ATT.	×	Custom	Sent.
	Tang et al. [91]	BERT [22]	×	×	×	ATT.	●	Custom	Sent.
	Alves et al. [6]	BERT-based [22, 84]	×	×	×	ATT.	×	Custom	Sent.
	Yan et al. [98]	BERT [22]	×	×	●	ATT.	●	Custom	Sent.
	Kim et al. [45]	None	●	×	●	ATT.	×	[82], Cust.	Doc.
	TTPXHunter [80]	SecureBERT [2]	●	×	×	ATT.	●	Custom	Sent.
	You et al. [100]	CTI-BERT [75]	×	●	●	ATT.	×	[82], Cust.	Sent.
	ALERT [77]	SciBERT [11]	×	×	×	ATT.	×	Custom	Sent.
	Li et al. [56]	BERT-based [22, 84]	×	×	●	ATT.	×	[82], Cust.	Sent.
	CTI-to-MITRE [72]	Multiple [60, 65]	×	×	●	ATT.	●	[82], Cust.	Both
	HMCAT [35]	SecureBERT [2]	●	×	×	ATT.	×	Custom	Both
	MITREtrieval [37]	RoBERTa [62]	×	×	●	ATT.	●	Custom	Sent.
	Fayyazi et al. [24]	BERT-based [2, 62]	×	×	●	ATT.	×	Custom	Sent.
Unlabeled	LADDER [5]	SBERT(MPNet) [81]	–	–	●	ATT.	×	Custom	Sent.
	Kumarasinghe et al. [49]	SentSecBERT [49]	–	–	●	ATT.	×	[82], Cust.	Sent.
	Abdeen et al. [1]	ATTACK-BERT [1]	–	–	●	ATT.	×	Custom	Sent.

Legend: (●) Element is present. (×) Element is not present or unclear from text. (–) Element does not apply. (Augment.) Data Augmentation. (Mult.) Multi-label. (Ont.) Ontology. (CTRs) TTP extraction from real-world CTI reports. (Gran.) Granularity. (Doc.) Document-level. (Sent.) Sentence-level.

Literature Trends - Classification

Table 2: Classification approaches for TTP extraction.

	Paper	Model	Augment. Art.	Augment. OOD	Mult.	Ont.	CTRs	Data	Gran.
Labeled	rcATT [52]	Word2Vec [65]	×	×	●	ATT.	●	Custom	Doc.
	Ayoade et al. [7]	None	×	×	●	ATT.	●	Custom	Doc.
	SeqMask [30], Ge et al. [31]	FastText [12]	×	×	●	ATT.	●	Custom	Doc.
	TRAM [82, 99]	SciBERT [11]	×	×	●	ATT.	●	[82]	Sent.
	Liu et al. [61]	Custom	×	×	●	ATT.	●	Custom	Doc.
	TTPHunter [79]	BERT-based [22]	×	×	×	ATT.	●	Custom	Sent.
	TIM [101]	SBERT [81]	×	×	×	ATT.	×	Custom	Sent.
	Tang et al. [91]	BERT [22]	×	×	×	ATT.	●	Custom	Sent.
	Alves et al. [6]	BERT-based [22, 84]	×	×	×	ATT.	×	Custom	Sent.
	Yan et al. [98]	BERT [22]	×	×	●	ATT.	●	Custom	Sent.
	Kim et al. [45]	None	●	×	●	ATT.	×	[82], Cust.	Doc.
	TTPXHunter [80]	SecureBERT [2]	●	×	×	ATT.	●	Custom	Sent.
	You et al. [100]	CTI-BERT [75]	×	●	●	ATT.	×	[82], Cust.	Sent.
	ALERT [77]	SciBERT [11]	×	×	×	ATT.	×	Custom	Sent.
	Li et al. [56]	BERT-based [22, 84]	×	×	●	ATT.	×	[82], Cust.	Sent.
	CTI-to-MITRE [72]	Multiple [60, 65]	×	×	●	ATT.	●	[82], Cust.	Both
	HMCAT [35]	SecureBERT [2]	●	×	×	ATT.	×	Custom	Both
	MITREtrieval [37]	RoBERTa [62]	×	×	●	ATT.	●	Custom	Sent.
	Fayyazi et al. [24]	BERT-based [2, 62]	×	×	●	ATT.	×	Custom	Sent.
Unlabeled	LADDER [5]	SBERT(MPNet) [81]	–	–	●	ATT.	×	Custom	Sent.
	Kumarasinghe et al. [49]	SentSecBERT [49]	–	–	●	ATT.	×	[82], Cust.	Sent.
	Abdeen et al. [1]	ATTACK-BERT [1]	–	–	●	ATT.	×	Custom	Sent.

Legend: (●) Element is present. (×) Element is not present or unclear from text. (–) Element does not apply. (Augment.) Data Augmentation. (Mult.) Multi-label. (Ont.) Ontology. (CTRs) TTP extraction from real-world CTI reports. (Gran.) Granularity. (Doc.) Document-level. (Sent.) Sentence-level.

Literature Trends - Generation

Table 3: Generative approaches for TTP extraction.

Title	LLM Model	Ontology	SFT	FSP	RAG	CTRs	Doc.	Data
aCTIon [88]	GPT-3.5$_{\text{Turbo}}$	ATT&CK	✕	✕	✕	●	✕	Cust.
CTINexus [15]	GPT-4	MALOnt	✕	●	●	●	●	Cust.
AECR [14]	ChatGLM3$_{\text{6B}}$	ATT&CK	●	✕	✕	●	✕	Cust.
Fayyazi et al. [24]	GPT-3.5	ATT&CK	✕	✕	●	✕	✕	Cust.
Fengrui et al. [102]	Llama2	ATT&CK	●	✕	✕	✕	✕	Cust.
IntelEX [97]	GPT-4o$_{\text{mini}}$	ATT&CK	✕	✕	●	●	●	Cust.
Kumarasinghe et al. [49]	Multiple	ATT&CK	✕	●	✕	✕	✕	Cust.
AttacKG+ [105]	GLM-4	ATT&CK	✕	✕	✕	●	✕	Cust.
Fieblinger et al. [26]	Multiple	Custom	●	●	✕	●	✕	Cust.

Legend: (●) = Method is present, (✕) = Method is not present, SFT = Supervised Fine-Tuning, FSP = Few-Shot Prompting, RAG = Retrieval-Augmented Generation, CTRs = TTP extraction from real-world CTI reports, Doc. = One-Shot inference of whole CTI report (in all stages). Data = dataset used for evaluation.

Literature Trends - Generation

Table 3: Generative approaches for TTP extraction.

Title	LLM Model	Ontology	SFT	FSP	RAG	CTRs	Doc.	Data
aCTIon [88]	GPT-3.5$_{Turbo}$	ATT&CK	×	×	×	●	×	Cust.
CTINexus [15]	GPT-4	MALOnt	×	●	●	●	●	Cust.
AECR [14]	ChatGLM3$_{6B}$	ATT&CK	●	×	×	●	×	Cust.
Fayyazi et al. [24]	GPT-3.5	ATT&CK	×	×	●	×	×	Cust.
Fengrui et al. [102]	Llama2	ATT&CK	●	×	×	×	×	Cust.
IntelEX [97]	GPT-4o$_{mini}$	ATT&CK	×	×	●	●	●	Cust.
Kumarasinghe et al. [49]	Multiple	ATT&CK	×	●	×	×	×	Cust.
AttacKG+ [105]	GLM-4	ATT&CK	×	×	×	●	×	Cust.
Fieblinger et al. [26]	Multiple	Custom	●	●	×	●	×	Cust.

Legend: (●) = Method is present, (×) = Method is not present, SFT = Supervised Fine-Tuning, FSP = Few-Shot Prompting, RAG = Retrieval-Augmented Generation, CTRs = TTP extraction from real-world CTI reports, Doc. = One-Shot inference of whole CTI report (in all stages). Data = dataset used for evaluation.

Literature Trends

Problem: Existing solutions are largely incomparable!

Literature Trends

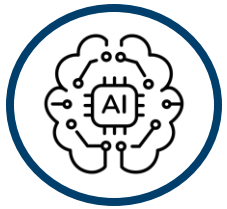
Problem: Existing solutions are largely incomparable!

PROBLEM

- **Custom datasets**
- **Not open-source**
- **Different TTP ontologies**
- **Handcrafted optimizations**

Empirical Study

Evaluation:
- Unified framework



Source Code:
- Re-Implementation of all 16 identified NLP components

Experiments:
- Lab Scenarios, Real-World Scenarios

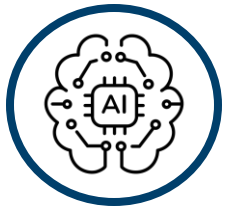
PROBLEM

- **Custom datasets**
- **Not open-source**
- **Different TTP ontologies**
- **Handcrafted optimizations**

Empirical Study

Evaluation:
- Unified framework

PROBLEM
- ~~Custom datasets~~
- Not open-source
- ~~Different TTP ontologies~~
- Handcrafted optimizations

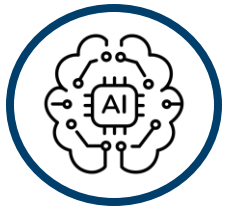


Source Code:
- Re-Implementation of all 16 identified NLP components

Experiments:
- Lab Scenarios, Real-World Scenarios

Empirical Study

Evaluation:
- Unified framework



Source Code:
- Re-Implementation of all 16 identified NLP components

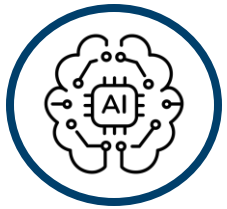
Experiments:
- Lab Scenarios, Real-World Scenarios

PROBLEM

- ~~Custom datasets~~
- ~~Not open-source~~
- ~~Different TTP ontologies~~
- Handcrafted optimizations

Empirical Study

Evaluation:
- Unified framework



Source Code:
- Re-Implementation of all 15 identified NLP components

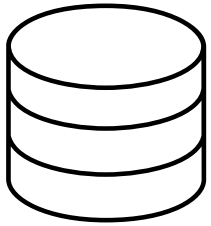
Experiments:
- Lab Scenarios, Real-World Scenarios

PROBLEM

- ~~Custom datasets~~
- ~~Not open-source~~
- ~~Different TTP ontologies~~
- ~~Handcrafted optimizations~~

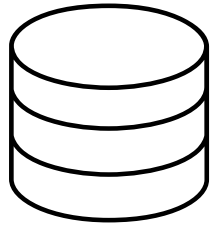
A Framework for Fair Comparison: Datasets

Datasets:
We used two public, real-world CTI datasets with sentence-level ATT&CK labels for training and testing.



MITRE TRAM2

150 reports
50 most common unique TTPs



Bosch AnnoCTR

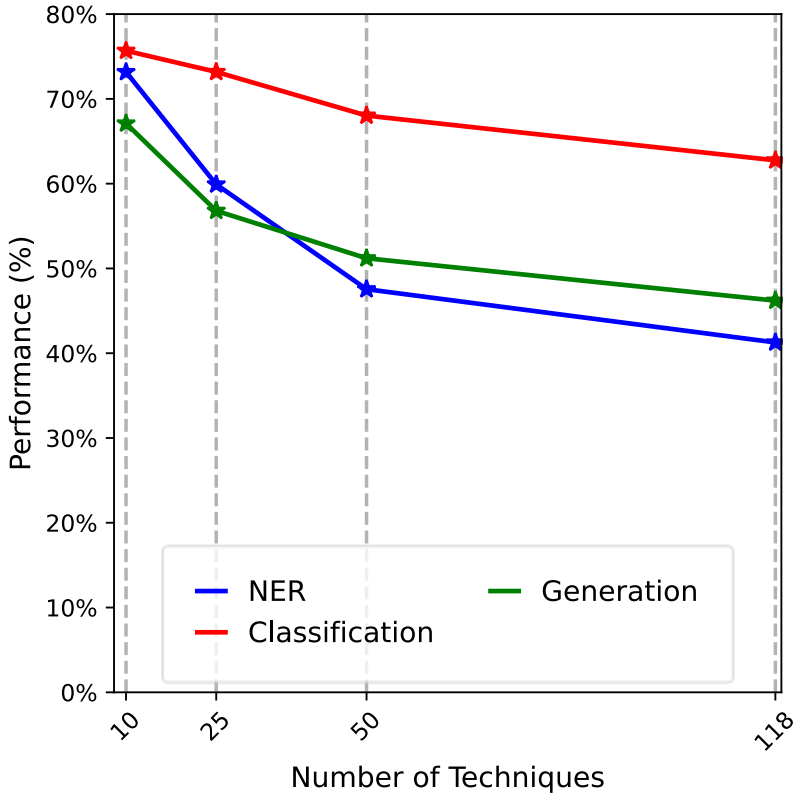
120 reports
118 unique TTPs

A Framework for Fair Comparison: Scenarios

Closed-Set: Simulates a lab environment where the set of possible TTPs is known beforehand. This is how most papers evaluate their work.

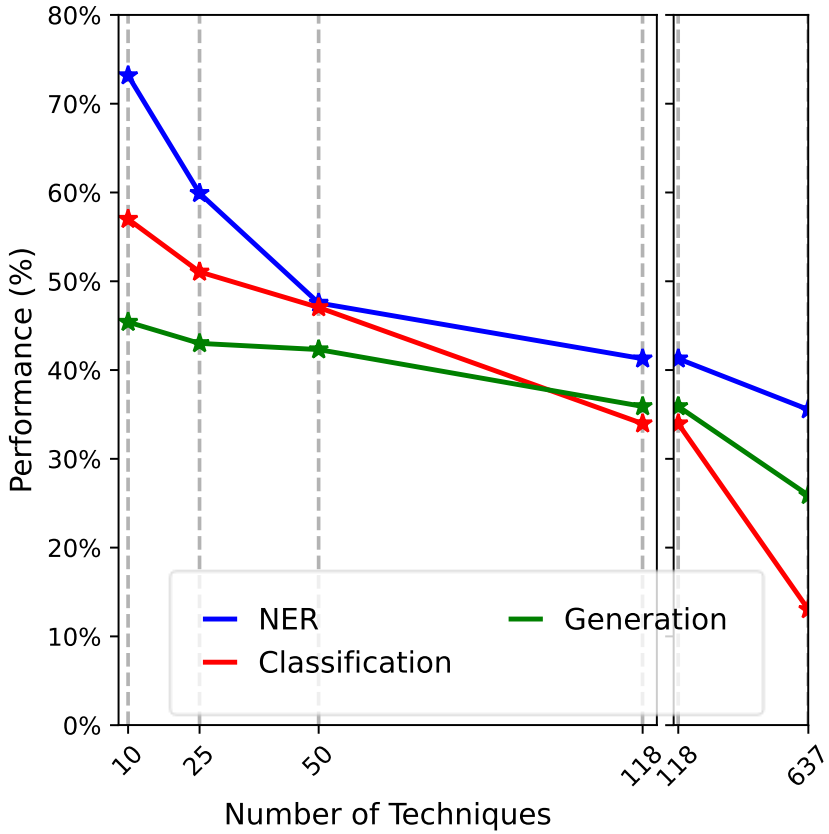
Open-Set: Simulates a realistic scenario where any TTP from the entire ATT&CK framework could appear. Models cannot be specifically tuned for the test data.

Key Finding 1: Closed-Set



In a controlled lab environment, modern, data-hungry models like LLMs and BERT-based classifiers are the top performers.

Key Finding 2: Open-Set



We tested on the AnnoCTR dataset against an increasing number of possible TTPs. This is where the results turn. The “older” traditional NER approach outperforms both newer data-driven approaches.

Key Finding 3: What are the bottlenecks?

The Dataset Bottleneck:

- Progress is fundamentally limited by the lack of large, high-quality, manually annotated public datasets.
- Existing datasets are small, cover only a fraction of ATT&CK techniques, and are highly imbalanced.
- Data augmentation techniques provide only marginal benefits.

Carl von Ossietzky Universität Oldenburg

Key Finding 3: What are the bottlenecks?

The Ontology Bottleneck (Label Confusion):

- The ATT&CK framework itself can be ambiguous. Many techniques are subtly different and hard to distinguish from a single sentence, even for a human expert.

- Example: T1547.001 (Boot or Logon Autostart Execution: Registry Run Keys) vs. T1112 (Modify Registry). Both involve modifying the registry, but the intent is different. Distinguishing them requires broader context.

Conclusion: Are We There Yet?

No. Automated TTP extraction is still an open research problem. Current approaches are not yet reliable enough for fully automatic practical use.

Conclusion: Are We There Yet?

The Path Forward is Not (Just) Better Models:

- The community's focus on novel NLP models is hitting a point of diminishing returns.
- Obstacles are the lack of high-quality data and the inherent ambiguities in the TTP definitions.

Recommendations:

- For Practitioners: A high-recall NER system can be a good assistant for analysts, while a fine-tuned classifier can work well for a known, closed set of TTPs.
- For Researchers: The community urgently needs to shift focus towards creating larger, high-quality, and more comprehensively annotated datasets.

Carl von Ossietzky
Universität
Oldenburg

Thank you!

Paper

uol.de/informatik/ssi

marvin.buechel@uol.de