Extracting Indicators of Compromise from Threat Reports by Leveraging the Power of LLMs

MAURICIO CROQUET THORNE, University of Twente, The Netherlands

Current advancements in cybersecurity often aim at facilitating the search and identification of vulnerabilities. These include Indicators of Compromise (IOCs), which serve as data artifacts that can be easily targeted/used with the intention of exploiting said systems. Cyber Threat Intelligence (CTI) reports are produced by cybersecurity teams with the expectation of exposing vulnerabilities in a given security framework. The act of exposing these vulnerabilities includes extracting IOCs. These indicators are commonly extracted from CTI reports by using what are known as rule-based extraction tools. However, these tools have limitations as they only extract known and correctly structured IOC types, causing them to overlook other potential indicators. Recent developments in Large Language Models (LLMs) suggest that these tools can be used to extract IOCs from CTI reports. This research explores the current differences between rule- and LLM-based extraction with expectations of understanding the difference in performance of both sides. This research showcases an LLMs ability to extract IOCs with high completeness (80% in average). Meanwhile, showing how rule-based extraction tools are consistent with better precision, as f1-scores outperform LLM-based extraction. Given that LLM-based tools showed to be more proficient at extracting IOCs from unstructured text while rule-based extraction showed consistency, a mix between both could yield promising results.

Additional Key Words and Phrases: Large language models, cyber threat intelligence, indicators of compromise, rule- and LLM-based extraction, prompt engineering

1 INTRODUCTION

In cybersecurity departments, Cyber Threat Intelligence (CTI) reports are often generated to describe any ongoing cyber threats. These reports tend to contain a mix of structured data within unstructured narrative/grammar. Due to to the lack of standardization in these reports, automated parsing of their data becomes a challenging task. A key aspect while analyzing CTI reports is the correct identification and extraction of Indicators of Compromise (IOC). IOCs are technical artifacts such as IP addresses, domain names, hash files, email addresses, among others which can impact the integrity of the given security framework. In addition, other artifacts such as Tactics, Techniques, and Procedures (TTP) are also obtainable from CTI reports. These go deeper than regular IOCs, as they describe how attackers operate.

The state-of-the-art rule-based extraction tools, such as IOCSearcher [16], depend on discrete and often limited methods. These tools are known for their consistency reproducing the same output when given the same input. They search for a limited and predetermined amount of IOC classifications within CTI reports. Consequently, the lack of standardization in these reports complicates the consistent extraction of IOCs [7] for rule-based extraction tools. These tools can miss important IOCs that are hidden in narrative text given that these may vary widely in structure.

The integration of Large Language Models (LLMs) into IOC extraction tools is still in early stages of development [11]. These tools are capable of acting in the same manner as rule-based extraction tools even though their efficacy and applications can be largely different. Through methods of prompt engineering it is possible to center an LLMs focus on a given task under useful context. LLMs have the potential to outperform conventional rule-based tools by identifying IOCs through contextual awareness, inferring threats, and adjusting to diverse linguistic patterns [7]. Nevertheless, the reliability of LLMs on this subject remains a question.

Diverse prompt engineering strategies can be used to alter performance of LLMs. By testing different strategies of prompt engineering it's possible to distinguish which of these can benefit LLMs the most while extracting IOCs. Three main methods of prompting can showcase the difference in IOC extraction performance. Namely, zero-, one- and few-shot prompting, each named relative to the amount of contextualization provided in their prompts. While zero-shot prompting provides no context or examples, one- and few-shot prompting offers realistic examples alongside their respective expected outcome.

Furthermore, this research also focuses on measuring the difference in performance between rule- and LLM-based extraction tools. This is done by extracting IOCs from CTI reports using both of these tools and then assessing them on their separate performance. Given that rule-based extraction tools are limited to extracting known IOC types, two separate metrics of performance will be considered. One will measure IOC extraction performance based on all known IOC types for said rule-based method. As it is possible some IOC types in the selected CTI reports are unknown to the selected rule-based extraction tool. The other metric will take into consideration all potential IOC types in the CTI reports.

Recent research has shown how hybrid approaches that combine deterministic tools (e.g. IOCSearcher) or structured formats (e.g., STIX [4]) with LLMs are gaining attention, notably IntelEX [10]. Approaches such as these aim to leverage the potential of each paradigm: the precision and consistency of rule-based tools and the versatility and contextual awareness of LLMs. The effectiveness of such integration strategies

TScIT 43, July 4, 2025, Enschede, The Netherlands

is yet to be evaluated, especially in terms of improving extraction quality and producing actionable CTI reports and TTPs. This research investigates the synergies of rule- and LLM-based IOC extraction methods.

Finally, the two research questions can be derived by what has been previously described:

- **RQ1**: To what extent is it possible to impact the performance of a LLM-based IOC extraction tool through different methods of prompt engineering?
- **RQ2**: How does the performance of a state-of-the-art rule-based tool (IOCSearcher) compare to LLMs in terms of accuracy, completeness, f1-score and adaptability when extracting IOCs from CTI reports?

2 BACKGROUND KNOWLEDGE

The extraction of IOCs within CTI reports has become an integral part of mitigating vulnerabilities in any given security setting [14]. This leads to the development of tools that automate the extraction of IOCs. Rule-based extraction tools have been employed to parse IOCs from reports using predefined patterns such as regular expressions or fixed token sequences [4], [16].

Multiple extraction tools have been developed making use of different novel technologies, notably rule-based [16], deep learning (DL) [26], machine learning (ML) [29], LLMs [1] and knowledge graph (KG) based [23] techniques. Each approach has its drawbacks, which impact what method can be chosen for this research. Rule-based systems lack contextual awareness [22], which limit their capacity to interpret indicators differently depending on their environment. DL and ML techniques require large datasets to train on, making it complicated to only filter valid information [6]. LLM-based extraction tools are limited to their context window and often experience "hallucinations" while extracting indicators [3]. Given that all the present methods pose limitations, manual labeling remains the most reliable method [16], [6].

Feeding incorrect ground truth (GT) data to LLMs seriously impacts their ability to correctly evaluate CTI reports [18]. It becomes necessary to guarantee a level of certainty in labeling indicators if the results are a consequence of them. This is why this research will base its GT strictly by using manual labeling to determine all possible IOCs that can be extracted from their respective CTIs.

Different prompting strategies can have an impact on the level of contextualization considered by the LLM [24]. Strategies such as zero-shot, one-shot, or few-shot learning can be applied to tailor model behavior without the need for explicit retraining [24]. The name of these strategies describes the level of contextualization and classification fed to the LLM through a prompt. Together, these evolving methods form the foundation for new approaches to IOC extraction, aiming to overcome the limitations of rule-based tools and better leverage the rich but variable language found in CTI reports.

3 RELATED WORKS

Automated rule-based IOC extraction has been a prominent choice for cybersecurity specialists since 2016 with the rise of some of the first rule-based tools such as iACE [27]. These tools work by searching for predefined patters, usually these consist of regular expressions (regex) or static heuristics. This allows the tools to identify and extract IOCs from unstructured text. However, these tools experience shortcomings such as a lack of contextual awareness or high false positive/negative rates. Regardless, the use of rule-based extraction tools is still important. Current relevant tools using this approach include: iACE, IoCMiner, IOCSearcher, among others.

Novel hybrid solutions have been recently developed with the expectation of benefiting from the strengths of LLM-based extraction tools. Notably, IntelEX [10] which uses LLMs to "purify" text, facilitating the extraction of their respective IOCs. This tool has shown it can outperform current state-of-the-art approaches such as AttackKG [28]. Additionally, “transformer-based models” are also used to identify structured threat intelligence from free text reports [17]. These models classify threat intelligence entities with the same expectations of this research, the only difference being; these models are solely based on LLMs.

Moreover, different AI approaches have undergone recent development, namely CTINexus [8]. This approach made use of ML and KGs to improve an LLMs capacity for in-context learning. The results of this research showed promising results for completeness and adaptability. However, the end of this research mentioned excessive fine tuning resulted in overfitting which compromised the extraction of these indicators. In addition, STIXnet [9] is another approach making use of alternative AI options. It uses a Natural Language Processing (NLP) framework extracting STIX objects and relations from CTI reports. This alternative experienced complications as the number of STIX classes increased, showing problems in scalability.

Lastly, in You et al. [13] researchers developed a model to extract Tactics, Techniques, and Procedures (TTPs) with positive accuracy readings (0.941). However, TTPs don’t encapsulate the spectrum of cybersecurity on its entirety.

4 METHODOLOGY AND APPROACH

This chapter details how each of the research questions (RQ) will be tested. Starting by defining the metrics and parameters needed to answer each RQ under controlled environments. Each RQ requires selecting extraction methods relative to their focus.

4.1 LLM Selection

Despite many AI approaches are suitable for this research, many of these can’t apply contextual prompting strategies, such as few-shot learning [2], [7]. LLMs offer versatility while interpreting and processing natural language. The selected

LLM will be used to test both research questions as both methods require an LLM for the same purpose. For this research, the minimal viable requirements for LLM-based IOC extraction protocols are the following:

- **Textual User Interface (TUI)**: A plain-text input alternative is necessary for inputting raw and complete CTI reports as well as prompt instructions into the model. A command line interface will facilitate such process as it takes solely textual responses.
- **Acceptable Token Range**: CTI reports often exceed the expected character limit LLMs have to accept prompts, the chosen LLM supports a sufficiently large token limit to process the complete prompt in one single interaction. Furthermore, the research also considers prompts that exceed the allowed token size to consider the disadvantages this brings.
- **Structured Output in JSON**: Structured Output in JSON: To allow programmatic comparison with other tools and support downstream automation, the model can return structured data in JSON format.

Following the criteria stated above and taking into consideration all available LLMs in the university's High Performance Computer (HPC) cluster, three main alternatives are present. These include Gemma3 [25], Deepseek-r1 [5] and Qwen2.5, all of which work as a suitable choice. Upon investigation, Qwen2.5 has been reported to outperform all conventional LLM models under dense information extraction [12]. Therefore, Qwen2.5 was selected. This LLM only works with text as input, and it supports 128 thousand tokens. According to its developers, Qwen2.5 [20] uses 1 token for every 3 - 4 characters on average while prompting in English. This translates to an average prompt size in the range of 32 to 42 thousand characters.

4.2 Rule-Based Extraction Tool Selection

Given that RQ2 compares rule- and LLM-based extraction, the selection of a deterministic rule-based extraction tool requires justification. IOCSearcher is specially designed for IOC extraction from unstructured CTI reports, this is a non-AI based model that uses regex with context scoring [16]. What stands out with IOCSearcher is the ability it has to handle noisy or unstructured text such as defanged expressions. Considering that, this tool can be seen as a state-of-the-art alternative. Adding on, IOCSearcher has been selected for the following reasons:

- Its capable of taking plain text input, allowing the same strategy to evaluate CTI reports with LLMs to be used.
- IOCSearcher is recognized as the most accurate tool on 11 out of the 13 IOC types supported by multiple tools [16]. The following is a reference of its known IOC types D.1.
- It outputs in `.iocs` or `.txt` formats, which can be easily stored and read to determine the difference between answers. While the prompt requests that the LLM answers in JSON format, IOCSearcher answers in plain text files and command line, as shown in the figure below.

```
(base) ➜  TREND CTI Reports iocsearcher -f Adware.Win32.ConduitS.txt -T
Searching into Adware.Win32.ConduitS.txt
email   engine@conduit.com
fqdn    ation.engine.conduit-services.com
fqdn    blockedration.engine.conduit-services.com
fqdn    conduit.com
fqdn    gle.com
fqdn    go.microsoft.com
fqdn    map.conduit-services.com
fqdn    ourtoolbar.com
fqdn    prefs.new
fqdn    s.conduit.com
fqdn    tmenu.engine.conduit-services.com
uuid    001709de-38a5-4f77-a69b-2f284030ddff
(base) ➜  TREND CTI Reports
```

Fig. 1. IOCSearcher IOC extraction

4.3 Dataset Selection

The data set used for this experiment was carefully evaluated and curated to ensure consistency across all CTI reports evaluated by limiting formatting variations. To make this possible, all reports have been collected from one single source [19], meant to minimize the differences in their structure and language. The length of these reports is under control, as it can impact on the performance and capabilities of the selected LLMs. All the reports were gathered in their original form, avoiding reports with images inside their technical details [D.2], as an attempt to preserve their authenticity to reflect real world conditions more accurately. 22 distinct CTI reports were selected, this sample size is large enough to notice the impact each research question has on IOC extraction. Furthermore, these CTI reports must include relevant IOC types that are known and unknown by the rule-based extraction tools selected.

4.4 Ground Truth and Manual Labeling

In order to evaluate the results of both RQs, a standardized method to classify IOCs in CTI reports has been established. By going through all CTI reports while manually labeling the present IOCs in each, a list of ground-truth extractions can be determined. The table below shows how established ground-truth can be compared to predicted extraction. This comparison can then determine the classification of each prediction. The potential outcomes are shown and detailed in the table below.

		Predicted	
		Positive (IOC)	Negative (No IOC)
Ground Truth	Positive (IOC extracted correctly)	True Positive (Ground truth IOC matches Predicted IOC)	False Negative (LLM missed manually labeled IOC)
	Negative (No IOC present)	False Positive (IOC provided while no ground truth IOC is present)	True Negative does not apply

Table 1. Predicted IOC Classification based on Ground Truth

4.4.1 Ground Truth classification . This section defines each potential IOC type encountered. These descriptions determine the validity of all IOCs present in the CTI reports used. The definitions below allow polymorphic representations of the

indicators to be considered. This will benefit the range of results recorded for the research. It is important to define these indicators openly as part of the purpose of this research is to investigate the impact contextualization and interpretation have in extracting IOCs. Here is a list of the IOCs and their possible interpretations:

- **IPv4 address** Accepted in regular expression format: 32 bit numerical address, separated by periods. Additional consideration of defanged alternatives (i.e. `BLOCKED.0.0.0`, `1[.]1.1.1`, etc)
- **Port Number** Unsigned 16 bit integer that acts as a unique identifier. Usually defanged as port under reports collected.
- **Fully Qualified Domain Name (FQDN)** Abbreviations of FQDNs (`https://www.google.com` to `google.com`) will be regarded as equal, both are accepted as valid active IOCs. Use of the "site" denomination is present in current CTI reports, this implies that the said fqdn has been defanged.
- **Filename(s)** Only requires said filename to be at the end of a filepath (if any) followed by file notation (such as `.txt`, `.pdf`, etc)
- **Filepath(s)** No methods to alter/defang filepaths has been noticed through the experimentation. Filepaths often display different representation relative to the operating system they use.
- **HashKey/files** These include hexadecimal or base64 representations of any given string with structured format such as: SHA256, MD5, BCRYPT, among others.
- **Username(s)** Defanged IOCs such as: USER, User name, etc, represent usernames in the selected CTI reports.
- **UUID** 32 hexadecimal characters grouped into five sections separated by hyphens.
- **Email address** Common email formatting is used, considering defanged emails do not exist in the report library in hand.

4.5 Evaluation Metrics

The evaluation of each RQ will be based on accuracy, completeness, f1-score and adaptability, using the manually annotated ground truth labels for each CTI report 1. Furthermore, the use for true positive (TP), false negative (FN) and false positive (FP) ground-truth can be found in the table below:

- **Accuracy** Accounts for correctly classified IOCs within all extracted IOCs.

$$\frac{\text{Correct IOC Classifications}}{\text{Total Classifications}} \equiv \frac{TP}{TP + FN + FP}$$

- **Completeness** measures successful extraction of relevant IOCs

$$\frac{\text{Number of Correct IOCs extracted}}{\text{Total Number of Ground Truth IOCs}}$$

- **F1-Score** measures the classification accuracy of the models.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{where,}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and,}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Adaptability** serves as a subjective metric to observe the versatility of a extraction tool.

4.6 Controlling the Environment for RQ1

To ensure that the research can answer this question reliably, it is necessary to formulate a controlled prompt that can be adjusted to fit multiple scenarios and context. Moreover, this section also focuses on the execution environment to guarantee that the LLM can constantly perform at the same level.

4.6.1 Prompt Design and Standardization . The main body of the prompt will be treated as a structured variable. A set of predefined prompt alternatives has been developed to reflect three common prompting strategies.

- **Zero-shot B.5**: Prompts that can only instruct the model. These cannot contextualize in any way what is expected from the LLM, meaning; this approach bases its answer solely on what indicators it is already aware of.
- **One-shot B.6**: Prompts that include a single example of CTI report format and IOC extraction, both with ground-truth and previously determined indicators.
- **Few-shot B.7**: Prompts that include two different examples before presenting the real CTI text for extraction.

Although each type differs in the degree of contextual support, all prompts share a common structure in terms of:

- Expected output format (JSON)
- Task description ("Extract all the Indicators of Compromise (IOCs) from the following cyber threat intelligence (CTI) report..")
- Text input format (plain CTI report excluding images). These reports will be similar in terms of average length, general structure, and language used.

This ensures that differences in performance between prompt types can be attributed to the presence or absence of examples. Using the same opening and body of text for each of the three prompts guarantees context in examples will play an important role in the quality of the answers.

4.6.2 Controlled input and execution environment . All prompts are sent to the same LLM instance running on a stable HPC node. This ensures equal processing power, runtime behavior, and avoids stochastic variation across different platforms. All translates to a system that will be evaluated mainly by how context changes the quality of the answers.

In addition, the same CTI report cluster will be used in every prompt strategy. Each report is processed independently using

zero-, one-, and few-shot prompts, these will be compared to manually extracted IOCs from these reports. Finally, the example CTI report with its respective answers will be used in both one- and few-shot approaches. This is meant to guarantee few-shot prompting will be given more contextualization.

4.7 Controlling the Environment for RQ2

This section considers both rule- and LLM-based extraction tools investigating their difference in performance. By selecting the best LLM-based strategy at extracting IOCs from CTI reports (in terms of accuracy, completeness and f1-score), both tools can be compared under optimal circumstances.

4.7.1 Evaluating rule-based IOC extraction under separate classification. Considering that IOCSearcher can only consider a limited amount of IOCs. This research question will consider two different outcomes. A comparison between rule- and LLM-based extraction tools were all potential IOC classifications are considered. And a comparison that only considers known IOCs for IOCSearcher. A list of all known IOCs for this rule-based extraction tool is attached in the appendix D.1.

4.7.2 Execution Consistency. Both systems are executed in a replicable environment:

- LLM prompts are processed using the same HPC infrastructure. In addition, prompt formatting ensures no significant changes in the task in hand.
- IOCSearcher is executed using a fixed command-line configuration with all CTI reports being sources from the same directory to ensure a fixed format in the text. (Example execution 1)

4.8 Experimental Setup

This section determines the tools used to compare the performance between rule- and LLM-based extraction while extracting IOCs from CTI reports. It outlines recommended tool selection with the intention of making this research replicable.

- Qwen2.5:latest
- IOCSearcher (2.5.10)
- Zero-, One- and Few-shot prompt templates
- Recommended GPU: 24GB VRAM (e.g., RTX 3090/4090)

4.9 RQ1 and RQ2 Methodologies

The following two subsections describe the steps required to replicate each of the two experiments.

4.9.1 RQ1 Methodology. Label ground-truth results for all available CTI reports C.1. The CTI source used included the following IOC types: IPv4 address, port number(s), FQDN, filename(s), filepath(s), hashkey(s), username(s), UUID, email address. Each report should consider each of these IOCs and report if they have them present. Formulate all necessary prompting strategies, these would be zero-shot prompting, one-shot prompting and few-shot prompting. Each should have the same initial template format and request. One-shot considers one CTI report and its results, meanwhile, few-shot considers that previous CTI report and a additional report showcasing most common scenarios. Prepare and run the necessary HPC scripts to have Qwen ready to receive prompts. Select one of the remaining prompting strategies to start recording results. Copy CTI report into the prompt structure. Paste and prompt/run said question in the given Qwen environment. Copy the answer to the LLM (it should come in a JSON format, given the request). Paste JSON object in results table and separate the string into different resulting IOCs extracted. Repeat these steps until no more CTI reports are left. At the end it is only necessary to compare the results to the manually labeled extraction and calculate the results.

4.9.2 RQ2 Methodology . To start the experiment, select the preferred prompting strategy Few-shot: B.7, One-shot B.6 or Zero-shot B.5 and collect labeled ground truth results from RQ1 test. Prepare and deploy all necessary HPC scripts, Ollama should be running Qwen2.5 and ready to take prompts. Copy each CTI report into the prompt structure, then paste and run said prompt to Qwen2.5. Copy the answer of the LLM (it should come in a JSON format, given the request). Paste the JSON object in the results table and separate the string into different resulting IOCs extracted. Then compare the results between the selected extraction strategy and the ground truth classification to calculate each individual metric. These steps should be recorded and repeated for all CTI reports gathered.

5 RESULTS

5.1 Quantitative Results

This section of the research is dedicated to the outcome of experimentation. Results of both research questions will be presented as information prior to interpretation.

RQ1 results: The table below displays the three chosen prompting methods used for LLM-based extraction along with the performance metrics achieved by each. These performance metrics represent the average result for each prompting strategy for 22 CTI reports.

	Average of Results (%)		
	Accuracy	Completeness	F1-score
Zero-shot	66	79	79
One-shot	72	80	85
Few-shot	72	81	84

Table 2. RQ1 Results

A diagram has been added to the appendix showing how an individual result can be obtained by comparing it with ground-truth labels C.2. Furthermore, an individual example of how one- and few-shot prompting extraction differs from zero-shot prompting can be seen in the figure below.

Fig. 2. Difference in extraction between prompting strategies

In this figure, we can see 4 methods extracting IOCs from the same CTI report. The first (leftmost) represents the ground truth with manual labeling. The second one represents Zero-shot prompting extraction, while the third and fourth example show one- and few-shot prompting respectively. In this specific example we see how accuracy jumps from 50% (from zero-shot prompting) to 100% (in both few- and one-shot prompting).

A final representation for results will be added to this section. The chart below displays the frequent accuracy readings from different prompting methods.

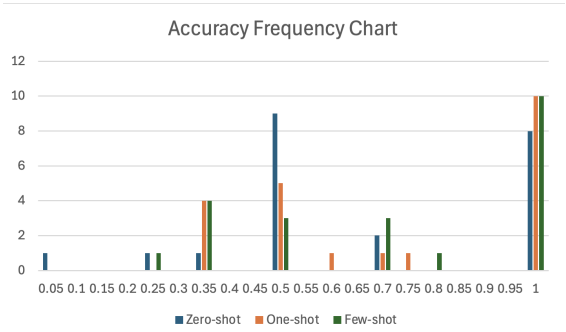


Fig. 3. Frequent Accuracy readings by different prompting strategies

RQ2 Results:

The following results show a comparison between the three metrics considered for RQ2 under each respective extraction method, even under specific consideration. For example, completeness for IOCSearcher with All IOCs Considered (AIC) is not recorded. This is because completeness accounts for all potential IOCs in the reference. This is impossible to consider for this metric as ground truth labeling considers a different range of indicators. Below, a key can be found to distinguish the two alternatives at measuring IOCSearcher performance.

Key	
OKR	Only Known IOCs
AIC	All IOCs Considered

Table 3. IOCSearcher Against CTI reports on IOCs

	Average of Results (%)		
	Accuracy	Completeness	F1-score
Few-shot	72	81	84
IOCSearcher OKI	80	77	87
IOCSearcher AIC	45	-	61

Table 4. RQ2 Results

In addition to these results, a scatter graph showcasing precision and recall recordings between IOCSearcher OKI and Few-shot prompting has been attached below:

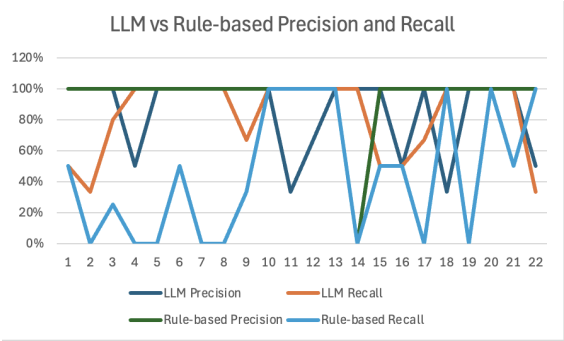


Fig. 4. LLM vs Rule-based Precision and Recall

5.2 Qualitative results

The only qualitative metric considered for this experiment occurs in RQ2. This metric has been labeled “adaptability”. The adaptability metric attempts to describe extraction versatility and correctness simultaneously. This metric can also be visualized with a table displaying the frequency of each adaptability label.

Missed Defanged Expression	Avoiding Unnecessary Extraction	Error in IOC Extraction	Complete Extraction of Expected IOCs
3 (14%)	2 (9%)	4 (18%)	13 (59%)

Table 5. IOCSearcher Adaptability

6 EVALUATION

This chapter assesses the performance and effectiveness of the proposed research method using metrics or criteria relevant to the research goals. This evaluation focuses on the performance metrics, manually labeled reports, the collection of results and the prompt formats used.

The performance metrics used showed to be suitable for the results gathered from the research. This takes in consideration the ground truth classification did not define True Negatives (TN). This classification was not necessary for any calculated metric so the results were not impacted by it. In the case TNs would be considered, the calculation of accuracy would have changed all results. This is because the formula commonly used for accuracy, which considers TN as part of the total number of extractions. This metric would have been impossible to account for, as there is potentially an infinite number of TNs.

Considering results derive from comparing their extraction to ground truth, the use of these manually labeled indicators

play a big role in the confidence we can attribute to the research. The validity of two individual IOCs found is open to interpretation, one would be the use of the IP: `127.0.0.1`. This specific IP Address is reserved for what is known as a loopback address. Loopback addresses are used to test IP software on the host computer while its not related to the computer's hardware. It was decided to consider this IP as a valid IOC as it holds the structure and definition of an IP Address previously stated. While IOCSearcher refused to extract it as a valid/malicious IP, all LLMs did the opposite.

The collection of results was a complicated task, as rule- and LLM-based extraction tools have different ways of providing output. While LLMs provided all IOCs found in JSON format, IOCSearcher returns extracted IOCs in plain text. This meant manual extraction of IOCs had to follow receiving JSON strings, adding a layer of work to the process. The use of JSON output could have contributed to hallucinations for the one- and few-shot LLM-based extraction tools. This is because LLMs might have seen exemplar IOC types within examples, later overfitting to find possibly incorrect IOCs.

Defanged expressions had an impact on RQ2 results. This is mainly due to the inability IOCSearcher showcased while attempting to extract incomplete/partial IOCs. In the 22 CTI reports analyzed, 4 contained defanged IPv4 addresses in a format IOCSearcher is unable to recognize. This had a relevant impact on the results of RQ2. Even if IOCSearcher can detect defanged IPv4 addresses under certain structures such as 1[.]1[.]1[.]1, the LLM-based extraction approach can adapt and interpret strings under any format. In this situation, defanged IPv4 expressions such as BLOCKED.BLOCKED.180.60, are undetectable by some rule-based extractors while LLM-based extractors had no problem in detecting all defanged addresses. Surprisingly, defanged expressions were identified throughout all prompting methods, showing how prompt context was not responsible for identifying those specific indicators.

RQ2 focuses on impact in performance between rule-based IOC extraction tools. The metrics evaluated in this section include the same three metrics mentioned, as well as an observation metric labeled adaptability. Adaptability observes how well LLMs adapt to different formats or limitations in CTI reports. These labels focus on the LLMs' ability to extract IOCs with defanged formats (meant to differ from their original IOC format), as well as account for unnecessary extractions and errors in the extractions. This metric shows how 49% of CTI reports evaluated by IOCSearcher have complications of any kind.

7 DISCUSSION

This chapter is meant to interpret the results of both research questions, as well as give observations around the behavior of both LLM and rule-based extraction methods. For this section, each RQ will be evaluated against its own results.

7.1 What do RQ1 results suggest

The results show accuracy and f1-score to be the metrics with most variability. While zero-shot prompting strategies only achieve 66% accuracy, few- and one-shot achieve 72%. Considering there is no change between accuracy amongst these two strategies, its possible LLMs reach a point of inflection in improvement provided by contextualization. More importantly, the f1-scores suggest a drop in precision from one- to few-shot strategies. Surprisingly, completeness does see a constant improvement for every strategy. Putting this into perspective, while one- and few-shot extraction had the same accuracy, few-shot prompting had better completeness (81% over 80%). This can be an example of how overfitting and hallucinations might have affected the final results. Additionally, Figure 3 shows a distribution curve (excluding 0% and 100% results). In this curve its possible to observe how higher accuracy readings where accumulated by few-shot prompting. While zero- and one-shot prompting distribute more frequently under a lower accuracy reading. F1-scores suggest LLMs start underperforming classifying IOCs accurately when presented too much context. Further observations can be made for this metric, as an f1-score consists of precision and recall, observing these two metrics separately may benefit the discussion. Results show that one-shot extraction had better recall than zero- and few-shot, while the opposite happened in respect to precision. This may be because less contextualization may lead to fewer false negatives, while more contextualization lead to fewer false positives.

7.2 What do RQ2 results suggest

These results suggest that IOCSearcher is capable of accurately extracting known IOCs while also presenting evidence of a lack of consideration of other potential IOCs in CTI reports. Furthermore, IOCSearcher presents a lack of adaptability for unknown cases, given that 41% of CTI reports evaluated had minor to significant extraction issues impacting its performance. Few-shot prompting accuracy fell between both IOCSearcher extraction methods. This demonstrates that CTI reports that consider an open amount of IOC classifications may benefit from using LLM-based extraction. Nevertheless, the highest accuracy reading came from IOCSearcher considering only known IOCs (OKI). This shows how rule-based extraction tools are more reliable extracting IOCs in general but not necessarily finding them. The completeness metric showed few-shot prompting provided more volume of correctly found IOCs than IOCSearcher. This can be due to overfitting as we see better f1-scores on OKI rule-based extraction.

8 LIMITATIONS

This chapter considers potential limitations that might jeopardize the overall quality of results in the experiment.

- Sample Size: The experiment sample size is the number of CTI reports presented to both the rule- and LLM-based extraction tools. These CTI reports come from the

same source. Given that the results show close margins between different prompting approaches, it could have been beneficial to consider larger sample sizes.
- Prompt Size: One key limitation in this experiment is the allowed prompt size that LLMs can interpret. LLMs such as Qwen limit the size of prompts they can interpret.
- Manual IOC classification: Throughout the experiment, results on both sides needed to be classified as true or false based on ground truth manually extracted from the CTI reports in question. This process included reading all CTI reports and manually extracting all known IOCs within them.
- Model Variability: The probabilistic nature of LLMs and repeated queries using the same prompt may result in different outcomes every time. While measures were taken to reduce this variability, such as initiating all prompts under the same pretense, it still introduces inconsistencies in its performance.

9 CONCLUSIONS

This research set out to compare the effectiveness of rule- and LLM-based extraction tools on their ability to correctly classify IOCs from CTI reports. It aimed to assess these tools' accuracy, completeness, adaptability and f1-scores with hopes to recognize strengths and weaknesses in both sides. Furthermore, this experiment also had a focus on evaluating different prompting strategies in optimizing LLM extraction performance.

RQ1: Prompt engineering showed a significant positive impact on the extraction of IOCs from CTI reports. Few-shot prompting achieved the best performance of all tested strategies; in terms of accuracy and completeness. One-shot prompting had the highest resulting f1-score, suggesting the amount of contextualization started to jeopardize the extraction tool. The use of contextualization showed notable and consistently positive results extracting IOCs. Moreover, this alternative displayed the ability of LLMs to recognize recommended patterns alongside a baseline of what they've been trained to do. This finding is significant because it shows how different LLMs might be able to achieve similar results regardless of their training. However, it is possible that more contextualization could've had a negative impact in few-shot prompting. Observing how accuracy stops incrementing after one-shot prompting, it is possible that an optimal amount of context might be ideal for LLM-based extraction. This could also mean that too much context might cause overfitting and hallucinations, leading to a unreliable model.

RQ2: Rule-base extraction tools have shown to be more reliable under known IOC types, while LLM-based extraction tools offer better versatility. Considering this, despite promising results, LLM-based alternatives stand no chance when it comes to familiar IOC type extraction against rule-based methods. This is clearly shown by how completeness for few-shot strategies is its best metric. Completeness describes a tools' capacity to find all potential IOC types, disregarding incorrect classifications. IOCSearcher OKI presented promising accuracy and f1-score metrics which translate to more precise extraction. Lastly, Figure 4 shows one of the main differences between rule- and LLM-based extraction. This graph shows how rule-based recall isn't as consistent as LLM-based recall while when it comes to precision, rule-based tools outperform LLMs by an evident margin. The difference between overall metrics between these two methods shows they could benefit from each other. Moreover, defanged expressions were only extracted by LLM-based tools even though IOCSearcher should be able to handle similar cases. This example shows how an LLMs versatility was more effective than a rule-based case exception.

10 FUTURE WORK

This research has shown that there is potential for hybrid implementations of IOC extraction tools between rule- and LLM-based approaches. One possible continuation of this research includes investigating the benefits of using each approach under the strengths they've shown.

- **LLM-based text purifier application**: The use of LLMs to purify unstructured text might have large benefits. This concept includes using the LLM to extract IOCs with defanged expressions as well as organizing textual structures. After which, rule-based extraction tools such as IOCSearcher could be used with the newly structured text to guarantee correct extraction of IOCs in the new CTI report. This approach shows that problems such as the extraction of incomplete data would be avoided. As the results suggest, there are IOCs that are only one of the two approaches can extract. Taking advantage of this situation, both tools might work in complementary to each other.
- **Automated ground truth generation**: Instead of basing ground truth on manual labeling, higher institutions can define what ground truth means. Taking cybersecurity frameworks such as the NIST [21] or ISO [15], one can determine with intrinsic validity what is labeled as ground truth.

11 ACKNOWLEDGMENTS

Gratitude is extended to this research's supervisor, Zsolt Kucsván, whose feedback and guidance helped shape the outcome of this research.

REFERENCES

[1] R. Harper A. Niakanlahiji, L. Safarnejad and B.-T. Chu. 2019. *Iocminer: Automatic extraction of indicators of compromise from twitter.* https://ieeexplore.ieee.org/document/9006562

[2] Lang H Kim Y Sontag D. Agrawal M, Hegselmann S. 2022. *Large Language Models are Few-Shot Clinical Information Extractors.* https://arxiv.org/pdf/2205.12689

[3] D. Mehta S. Pasquali B. Sarmah, T. Zhu. 2023. *Towards reducing hallucination in extracting information from financial reports using Large Language Models.* https://arxiv.org/pdf/2310.10760

[4] Sean Barnum. 2014. *Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX).* https://stixproject.github.io/getting-started/whitepaper/

[5] DeepSeek-AI and contributing authors. 2025. *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.* https://arxiv.org/pdf/2501.12948

[6] S. T. Frankum R. Perdisci M. Antonakakis1 A. Keromytis E. Froudakis, A. Avgetidis. 2025. *Uncovering Reliable Indicators: Improving IoC Extraction from Threat Reports.* https://arxiv.org/html/2506.11325v1

[7] Brown et al. 2020. *GPT-3 and few-shot learning potential. Language Models are Few-Shot Learners.* https://arxiv.org/abs/2005.14165

[8] Cheng et al. 2024. *CTINexus: Automatic Cyber Threat Intelligence Knowledge Graph Construction Using Large Language Models.* https://arxiv.org/html/2410.21060v2

[9] Marchiori et al. 2023. *STIXnet: A Novel and Modular Solution for Extracting All STIX Objects in CTI Reports.* https://arxiv.org/pdf/2303.09999

[10] Ming Zu et al. 2024. *IntelEX: A LLM-driven Attack-level Threat Intelligence Extraction Framework.* https://arxiv.org/html/2412.10872v1

[11] Xu et al. 2024. *Large Language Models for Cyber Security: A Systematic Literature Review.* https://www.sciencedirect.com/science/article/abs/pii/S0167739X23000535?via%3Dihub

[12] X. Y. C. Zhang et al. 2025. *CaseReportBench: An LLM Benchmark Dataset for Dense Information Extraction in Clinical Case Reports.* https://arxiv.org/html/2505.17265v1

[13] You et al. 2022. *TIM: threat context-enhanced TTP intelligence mining on unstructured threat data.* https://cybersecurity.springeropen.com/articles/10.1186/s42400-021-00106-5

[14] N. V. Verde F. Marchiori, M. Conti. 2023. *STIXnet: A Novel and Modular Solution for Extracting All STIX Objects in CTI Reports.* https://arxiv.org/pdf/2303.09999

[15] International Organization for Standardization. 2022. *ISO/IEC 27001: Information technology — Security techniques — Information security management systems — Requirements.* https://www.iso.org/home.html

[16] GoodFATR 2023. *The Rise of GoodFATR: A Novel Accuracy Comparison Methodology for Indicator Extraction Tools.* https://www.sciencedirect.com/science/article/abs/pii/S0167739X23000535?via%3Dihub

[17] et al. Husari, G. 2017. *TTPDrill: Automatic extraction of threat actions from unstructured text of CTI reports.* https://dl.acm.org/doi/10.1145/3134600.3134646

[18] N. Ihde A. Nathansen N. Noack H. Patzlaff F. Naumann L. Budach, M. Feuerpfeil and H. Harmouch. 2025. *The Effects of Data Quality on Machine Learning Performance on Tabular Data.* https://arxiv.org/pdf/2207.14529

[19] Trend Micro. n.d. *Threat encyclopedia: Malware.* https://www.trendmicro.com/vinfo/us/threat-encyclopedia/malware

[20] Qwen Team. (n.d.). 2025. *Byte-level Byte Pair Encoding.* https://qwen.readthedocs.io/en/v2.5/getting_started/concepts.html

[21] National Institute of Standards and Technology. 2018. *Framework for Improving Critical Infrastructure Cybersecurity.* https://www.nist.gov/

[22] Wenjie Dong Lin Ai Ziwei Gong Songfang Huang Zongsheng Li Ehsan Hoque Julia Hirschberg Yue Zhang Pai Liu, Wenyang Gao. 2024. *A Survey on Open Information Extraction from Rule-based Model to Large Language Model.* https://arxiv.org/html/2208.08690v6

[23] Nidhi Rastogi Romy Fieblinger, Md Tanvirul Alam. 2024. *Actionable Cyber Threat Intelligence using Knowledge Graphs and Large Language Models.* https://arxiv.org/html/2407.02528v1

[24] J. Zhu T. Xiao. 2024. *Efficient Prompting Methods for Large Language Models: A Survey.* https://arxiv.org/html/2404.01077v1

[25] Gemma Team. 2025. *Gemma3 Technical Report.* https://arxiv.org/pdf/2503.19786

[26] X. Wang Z. Li L. Xing X. Liao, K. Yuan and R. Beyah. 2016. *Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence.* https://doi.org/10.1145/2976749.2978315

[27] X. Wang Z. Li L. Xing X. Liao, K. Yuan and R. Beyah. 2016. *Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence.* https://dl.acm.org/doi/pdf/10.1145/2976749.2978315

[28] Yan Chen Zhenkai Liang Zhenyuan Li, Jun Zeng. 2021. *AttacKG: Constructing Technique Knowledge Graph from Cyber Threat Intelligence Reports.* https://arxiv.org/abs/2111.07093

[29] Z. Zhu and T. Dumitras. 2018. *Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports.* https://ieeexplore.ieee.org/document/8406617

12 APPENDICES

A AI USE DISCLOSURE

During the preparation of this research, ChatGPT was used to correct structural errors and improve flow of this research. After using this tool, the content was reviewed and edited as needed while taking full responsibility for the work.

B PROMPT TEMPLATES

```
Prompt: "Extract all the Indicators
of Compromise (IOCs) from the
following cyber threat intelligence
(CTI) report.
Return the output in JSON format
using the following structure:
{
"IOC-1": [],
"IOC-2": [],
"IOC-3": [],
"IOC-4": [],
"IOC-5": [],
....
}

CTI Report:
[CTI REPORT]|
```

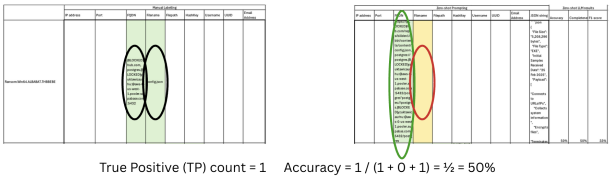
Fig. B.5. Zero-shot prompt used

C RESULT EXTRACTION

D IOCSEARCHER KNOWN IOCS REFERENCE

Filename	Manual Labelling								
	IP address	Port	FQDN	Filename	Filepath	HashKey	Username	UUID	Email Address
Ransom.Win64.ALBABAT.THBBEBE			{BLOCKED}hub.com, postgres.{BLOCKED}yzuktawicaurhu:@aws-0-us-west-1.pooler.supabase.com:5432	config.json					

Fig. C.1. RQ1 Manual Labeling extraction Example



True Positive (TP) count = 1
False Positive (FP) count = 0
False Negative (FN) count = 1

Accuracy = 1 / (1 + 0 + 1) = ½ = 50%

Fig. C.2. RQ1 Single Accuracy result Calculation

- URLs (url)
- Domain names (fqdn)
- IP addresses (ip4, ip6)
- IP subnets (ip4Net)
- Hashes (md5, sha1, sha256)
- Email addresses (email)
- Blockchain addresses (bitcoin, bitcoincash, cardano, dashcoin, dogecoin, ethereum, litecoin, monero, ripple, solana, stellar, tezos, tron, zcash)
- Phone numbers (phoneNumber)
- Copyright strings (copyright)
- CVE vulnerability identifiers (cve)
- Tor v3 addresses (onionAddress)
- Social network handles (facebookHandle, githubHandle, instagramHandle, linkedinHandle, pinterestHandle, telegramHandle, twitterHandle, whatsappHandle, youtubeHandle, youtubeChannel)
- Advertisement/analytics identifiers (googleAdsense, googleAnalytics, googleTagManager)
- Payment addresses (webmoney)
- Chinese Internet Content Provider licenses (icp)
- Bank account numbers (iban)
- Trademarks (trademark)
- Universal unique identifiers (uuid)
- Android package name (packageName)
- MITRE ATT&CK Technique identifiers (ttp)
- Spanish NIF identifiers (nif)
- TOX identifiers (tox)
- Amazon Resource Names (arn)

Fig. D.1. List of known IOCs by IOCSearcher

It sets the system's desktop wallpaper to the following image:

- {LOCKBIT_wallpaper}.png

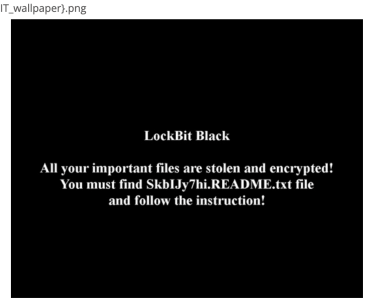


Fig. D.2. Image in Omitted CTI report

```
Prompt: "Extract all the Indicators
of Compromise (IOCs) from the
following cyber threat intelligence
(CTI) report.
Return the output in JSON format
using the following structure:
{
"IOC-1": [],
"IOC-2": [],
"IOC-3": [],
"IOC-4": [],
"IOC-5": [],
....
}

Example:
[CTI REPORT SECTION]

IOC Extraction:
[EXAMPLE IOC EXTRACTION ON CTI REPORT]

Actual CTI Report:
[CTI REPORT]
```

Fig. B.6. One-shot prompt used

```
Prompt: "Extract all the Indicators of
Compromise (IOCs) from the
following cyber threat intelligence (CTI)
report.
Return the output in JSON format using the
following structure:
{
"IOC-1": [],
"IOC-2": [],
"IOC-3": [],
"IOC-4": [],
"IOC-5": [],
....
}

Example 1:
[CTI REPORT SECTION]

IOC Extraction:
[EXAMPLE IOC EXTRACTION ON CTI REPORT]

Example 2:
[CTI REPORT SECTION]

IOC Extraction:
[EXAMPLE IOC EXTRACTION ON CTI REPORT]

Actual CTI Report:
[CTI REPORT]
```

Fig. B.7. Few-shot prompt used