Labeling NIDS Rules with MITRE ATT&CK Techniques using ChatGPT

Nir Daniel[1,2], Florian Klaus Kaiser[3,4], Anton Dzega[1,2], Aviad Elyashar[2,5], Rami Puzis[1,2]

[1] Department of Software and Information Systems Engineering, Ben-Gurion University, Israel

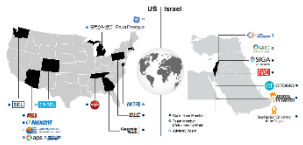
[2] Cyber@BGU, Cyber Labs at Ben-Gurion University

[3] Institute for Industrial Production, Karlsruhe Institute of Technology, Germany

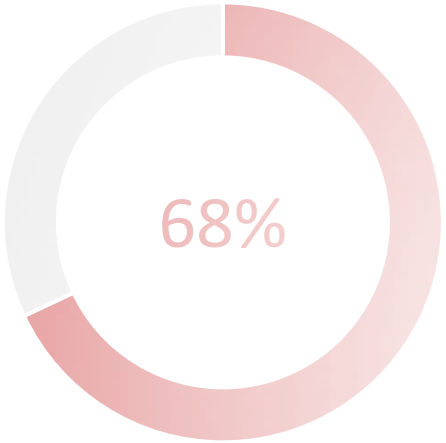
[4] Institute of Information Security and Dependability, Karlsruhe Institute of Technology, Germany

[5] Department of Computer Science, Shamoon College of Engineering, Israel

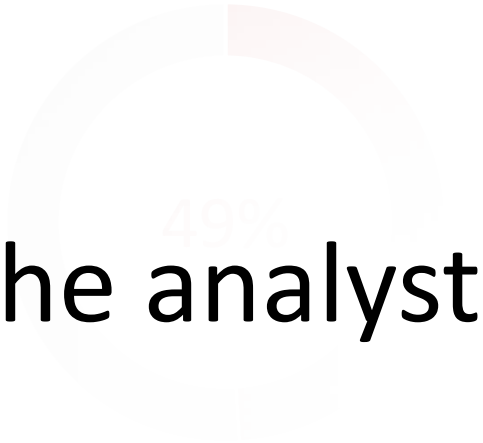
Introduction



According to a survey[1] conducted in 2020:

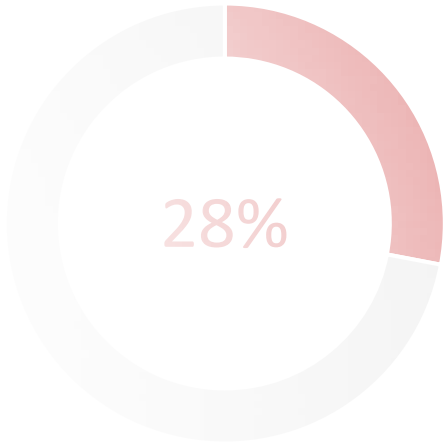


Reduce the alert volume of specific features



Turn off high-volume alerting features

Hire more analysts

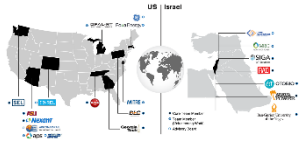


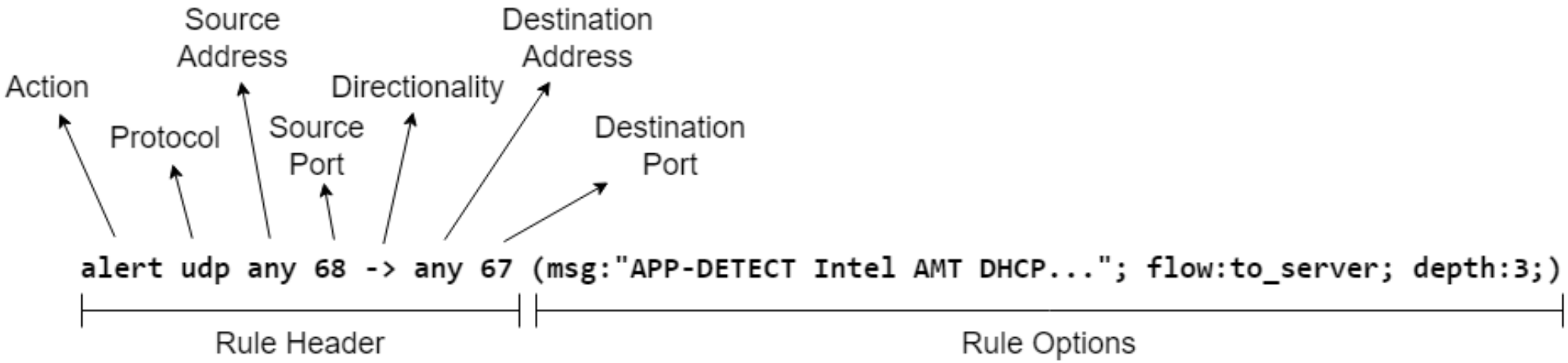
Ignore certain alert categories

The analyst is **overloaded**

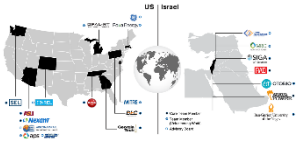
[1] Alert Overload Still Plagues Cybersecurity Industry – Critical Start

SNORT – Network Intrusion Detection System (NIDS)





MITRE ATT&CK



ATT&CK®

27 Tactics

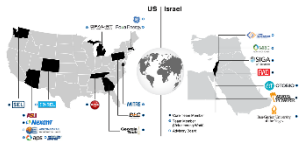
Reconnaissance . . . Impact

193 Techniques

Active Scanning . . . Phishing for Information

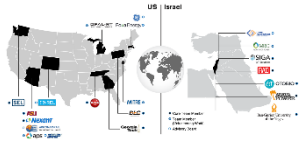
Disk Wipe . . . Firmware Corruption

Goal



Label NIDS rules with MITRE ATT&CK **techniques**

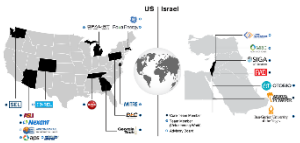
ChatGPT Performs Experts Tasks

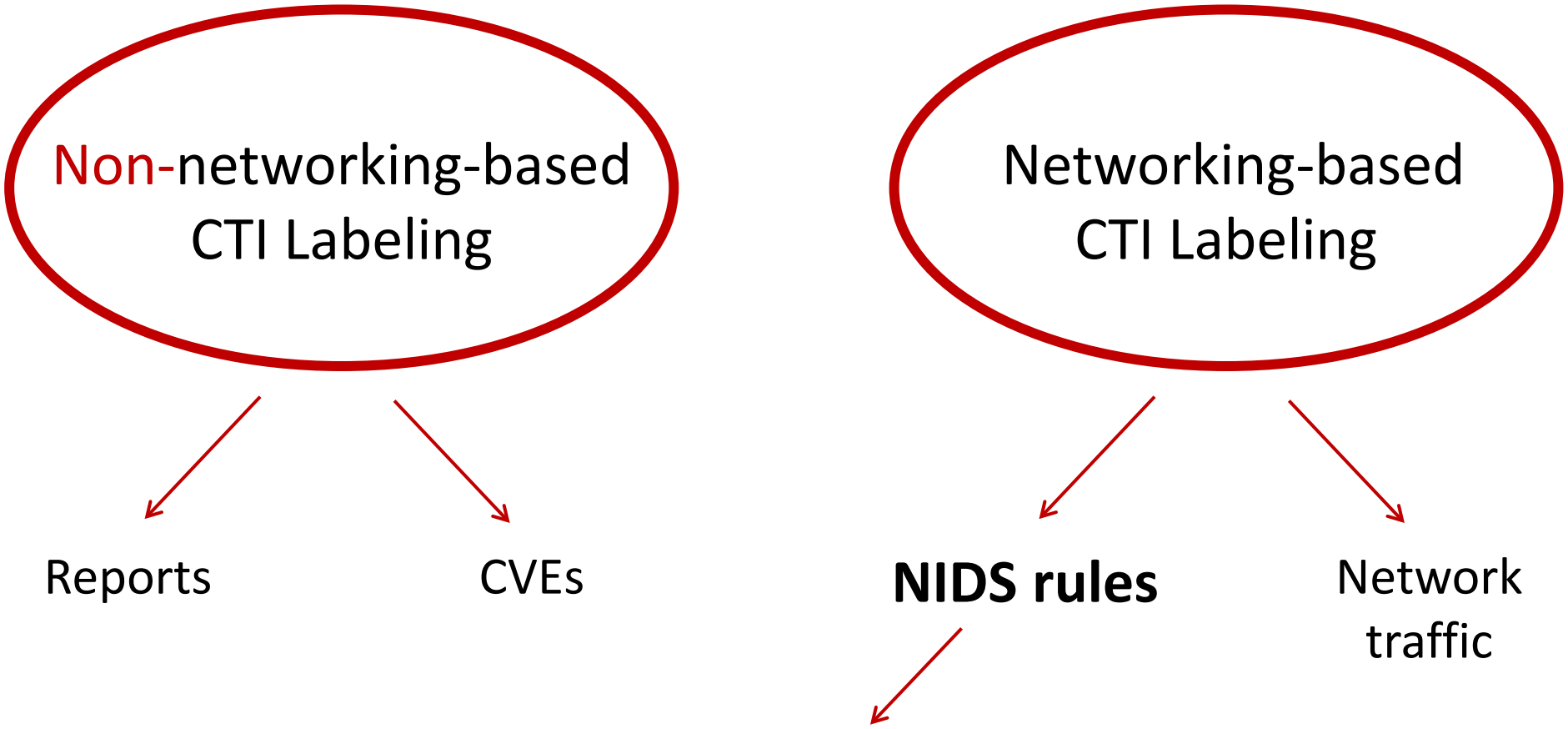


- ChatGPT for annotating political Twitter messages (Tornberg)
- ChatGPT passes the Canadian Head and Neck Surgery Examination (Long et al.)
- ChatGPT in cybersecurity for offensive actions (Tod-Raileanu et al.)

Idea: Why not use ChatGPT for labeling NIDS rules?

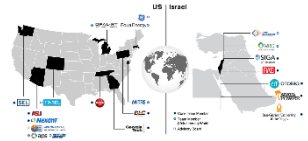
Related Work





Lin et al. – Attack tactic labeling for cyber threat hunting

Labeling NIDS Rules with MITRE ATT&CK Techniques



Rules

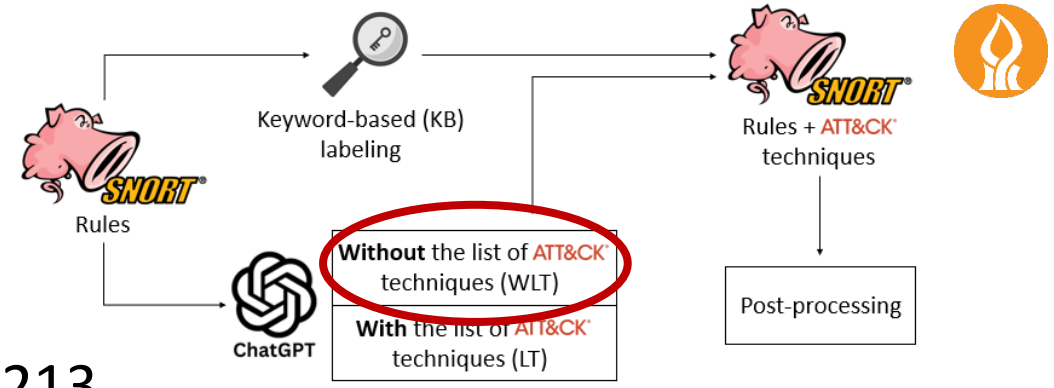
Keyword-based (KB) labeling

Rules + ATT&CK® techniques

Without the list of ATT&CK® techniques (WLT)
With the list of ATT&CK® techniques (LT)

Post-processing

Without the List of ATT&CK Techniques (WLT)



Snort Rule

alert tcp $EXTERNAL_NET any -> $TELNET_SERVERS 23 (msg:"MALWARE-BACKDOOR MISC Linux rootkit attempt"; flow:to_server,established; content:"wh00t!"; metadata:ruleset community; classtype:attempted-admin; sid:213; rev:9;)

ChatGPT

sid: 213,
Technique id: "T1210",
Technique name: "Exploitation of Remote Services",
Quotes: ["alert tcp $EXTERNAL_NET any -> $TELNET_SERVERS 23”],
Explanation: "The rule suggests exploitation of the Telnet service, a remote service. This corresponds to the 'Exploitation of Remote Services' technique.”

With the List of ATT&CK Techniques (LT)

Snort Rule

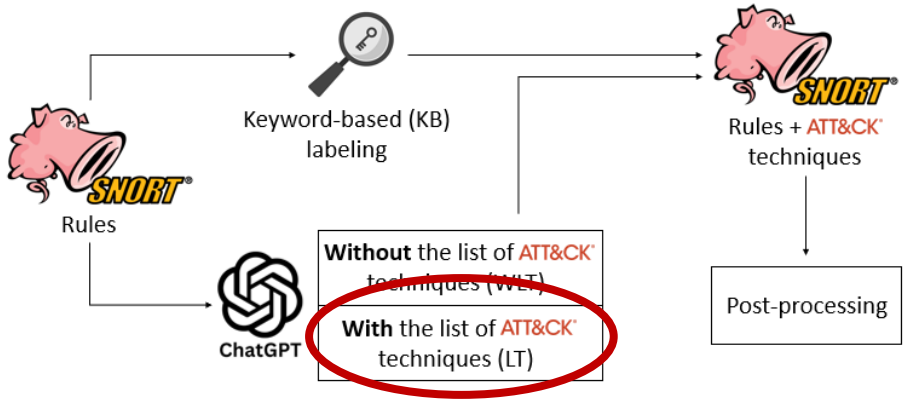
alert tcp $EXTERNAL_NET any -> $TELNET_SERVERS 23 (msg:"MALWARE-BACKDOOR MISC Linux rootkit attempt"; flow: to_server, established; content: "wh00t!"; metadata: ruleset community; classtype: attempted-admin; sid:213; rev:9;)

+

Full List of ATT&CK Techniques

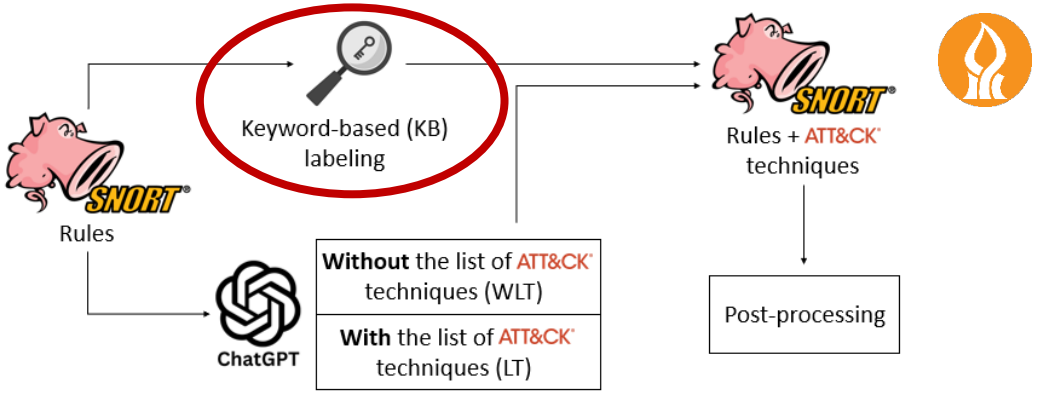
T1548 Abuse Elevation Control Mechanism
T1134 Access Token Manipulation
. . .

ChatGPT



sid: 213,
Technique id: "T1548",
Technique name: "Abuse Elevation Control Mechanism",
Quotes: ["MALWARE-BACKDOOR MISC Linux rootkit attempt"],
Explanation: "The rule implies an attempt to install a rootkit on the targeted Linux system. This …"

Keyword-based (KB) Labeling



Snort Rule

alert tcp $EXTERNAL_NET any -> $TELNET_SERVERS 23 (msg:"MALWARE-BACKDOOR MISC Linux **rootkit** attempt"; flow:to_server,established; content:"wh00t!"; metadata:ruleset community; classtype:attempted-admin; sid:213; rev:9;)

ATT&CK®

Technique T1014
“Rootkit”

Post Processing

Rules

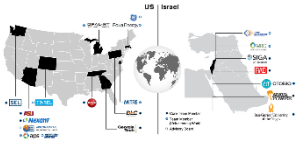
Keyword-based (KB) labeling

Without the list of ATT&CK® techniques (WLT)
With the list of ATT&CK® techniques (LT)

Rules + ATT&CK® techniques

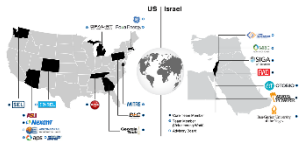
Post-processing

Evaluation



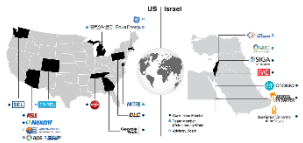
- A set of 162 labeled Snort rules from the official Snort repository
- An average of 1.38 technique labels per rule
- 30 unique techniques in the entire evaluation set

Experimental Setup



- Tested both **ChatGPT-3.5** and **ChatGPT-4**
- Questioned ChatGPT with **(LT)** and without **(WLT)** the list of ATT&CK techniques on each rule
- Applied the Keyword-based **(KB)** labeling on each rule
- Evaluated different combinations of methods
- **Metrics: Average Precision, Recall and F1-score values**

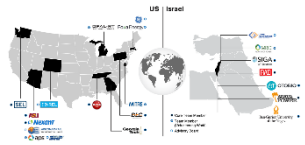
Experimental Setup – Frequency-based Baseline

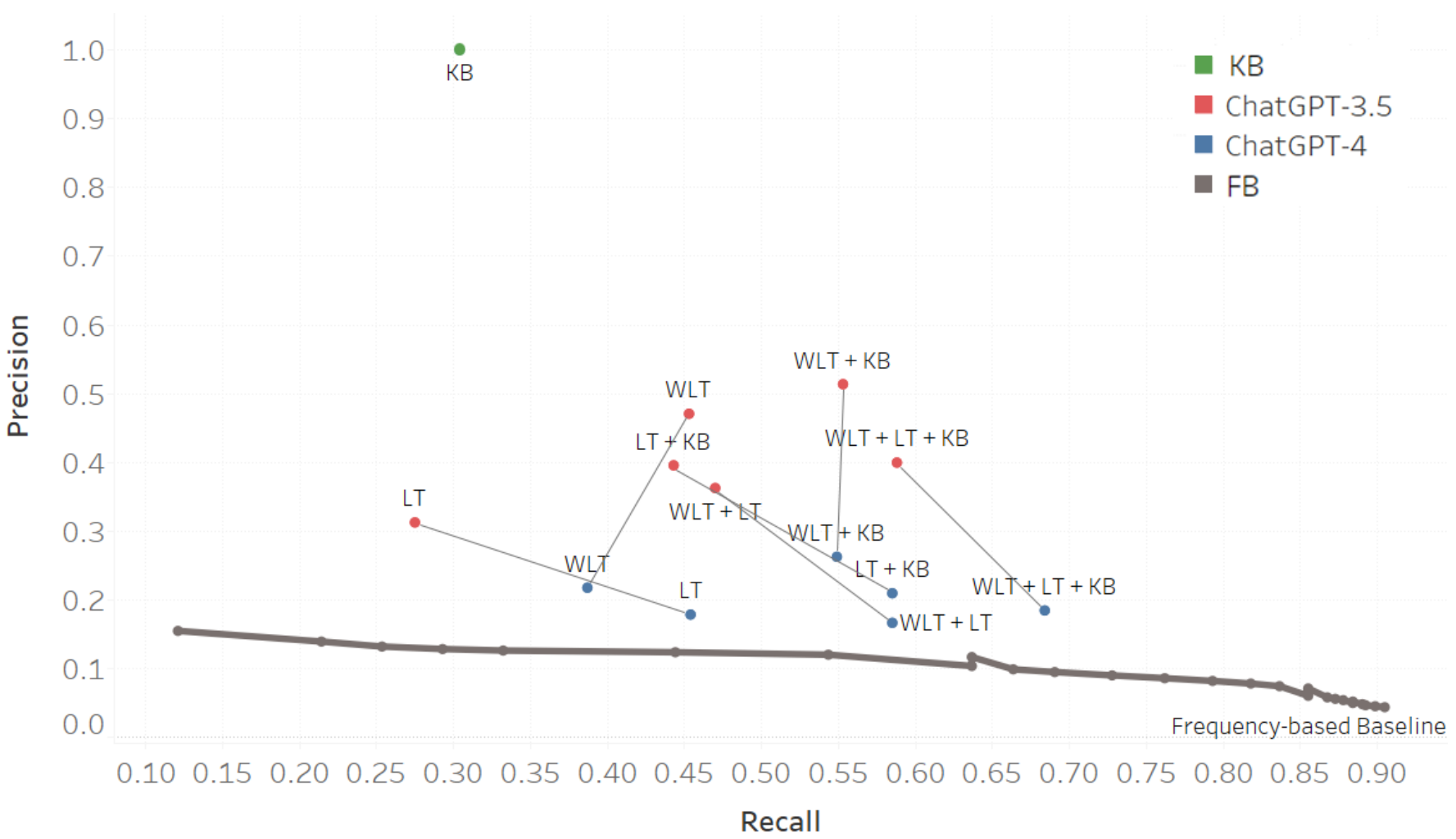


What will be my score if I will always select the *n*-most frequent techniques?

The Frequency-based (FB) Baseline measures the metrics for every *n*

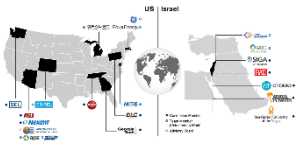
Results





Precision and Recall of every method

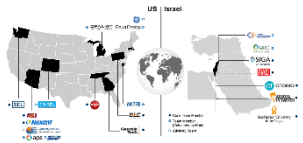
Results



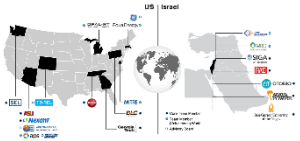
Model	Method	Precision	Recall	F1-score
	FB baseline	0.117	0.637	0.191
	KB	1	0.304	0.304
ChatGPT-3.5	WLT	0.471	0.453	0.433
	WLT + KB	**0.514**	0.553	**0.492**
	LT	0.313	0.275	0.285
	LT + KB	0.396	0.443	0.397
	WLT + LT	0.363	0.47	0.382
	WLT + LT + KB	0.4	0.588	0.437
ChatGPT-4	WLT	0.218	0.387	0.247
	WLT + KB	0.263	0.549	0.317
	LT	0.179	0.454	0.241
	LT + KB	0.21	0.585	0.29
	WLT + LT	0.167	0.585	0.241
	WLT + LT + KB	0.185	**0.684**	0.271

Average Precision, Recall and F1-score of each method

Conclusions



- Providing ChatGPT-3.5 with the list of techniques **weakened** the results, in contrast to ChatGPT-4
- It is always beneficial to combine ChatGPT with the Keyword-based method
- Interestingly, **ChatGPT-3.5 achieved better results** than ChatGPT-4
- We proposed a Proof-of-Concept of employing a publicly accessible GPT for labeling NIDS rules



Thank You!

Any Questions?

Contact: nirdanie@post.bgu.ac.il