CTIBench: A Benchmark for Evaluating LLMs in Cyber Threat Intelligence

Md Tanvirul Alam[1], Dipkamal Bhusal[1], Le Nguyen[1], and Nidhi Rastogi[1]

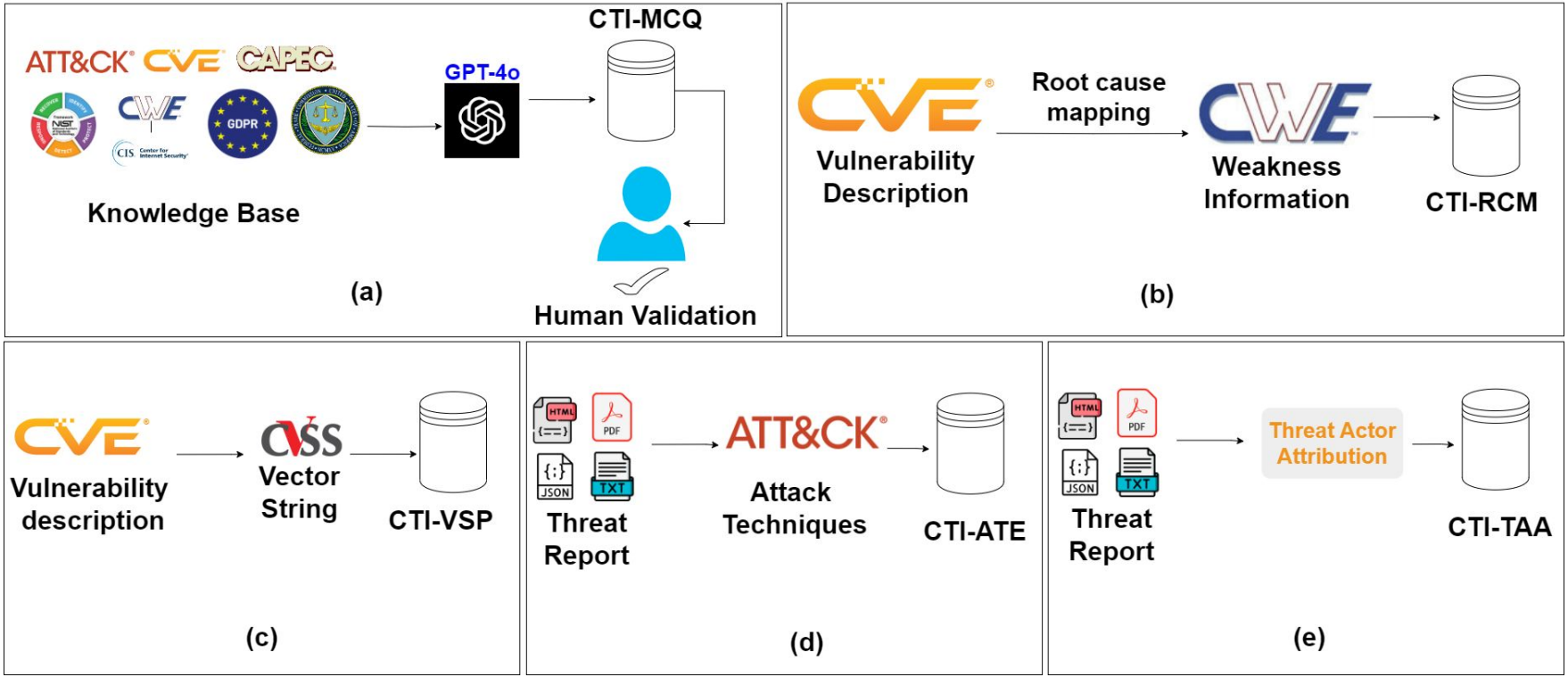
[1]Rochester Institute of Technology, Rochester NY 14623, USA

NeurIPS, 2024

Motivation: Why CTIBench?

- LLMs can transform CTI by processing vast amounts of unstructured threat data, but their tendency to hallucinate and misunderstand technical content poses risks for generating reliable intelligence.
- General benchmarks, like GLUE, fail to assess the practical challenges of cybersecurity, making it hard to measure LLM performance in CTI.
- Existing CTI benchmarks focus mainly on memorization, neglecting critical aspects like comprehension and problem-solving in real-world CTI scenarios.

CTIBench


ATT&CK
CVE
CAPEC
CWE
GDPR
CIS
Knowledge Base
GPT-4o
CTI-MCQ
Human Validation
(a)
CVE
Vulnerability Description
Root cause mapping
CWE
Weakness Information
CTI-RCM
(b)
CVE
Vulnerability description
CVSS
Vector String
CTI-VSP
(c)
HTML
PDF
JSON
TXT
Threat Report
ATT&CK
Attack Techniques
CTI-ATE
(d)
HTML
PDF
JSON
TXT
Threat Report
Threat Actor Attribution
CTI-TAA
(e)

Example: CTI-RCM

Prompt: Analyze the following CVE description and map it to the appropriate CWE. Provide a brief justification for your choice.

Description: In the Linux kernel through 6.7.1, there is a use-after-free in cec_queue_msg_fh, related to drivers/media/cec/core/cec-adap.c and drivers/media/cec/core/cec-api.c.

Correct Answer: CWE-416 (Use After Free)

Example: CTI-VSP

Prompt: Analyze the following CVE description and calculate the CVSS v3.1 Base Score. Determine the values for each base metric: AV, AC, PR, UI, S, C, I, and A. Summarize each metric's value and provide the final CVSS v3.1 vector string.

Description: In the Linux kernel through 6.7.1, there is a use-after-free in cec_queue_msg_fh, related to drivers/media/cec/core/cec-adap.c and drivers/media/cec/core/cec-api.c.

Correct Answer: CVSS:3.1/AV:L/AC:L/PR:L/UI:N/S:U/C:N/I:N/A:H
CVSS Score: 5.5

Example: CTI-ATE

Prompt: Extract all MITRE Enterprise attack patterns from the following text and map them to their corresponding MITRE technique IDs. Provide reasoning for each identification.

Description: adbupd is a backdoor utilized by PLATINUM, bearing similarities to Dipsind. It has the capability to execute a copy of cmd.exe and includes the OpenSSL library to encrypt its command and control (C2) traffic. Additionally, adbupd can achieve persistence by leveraging a WMI script.

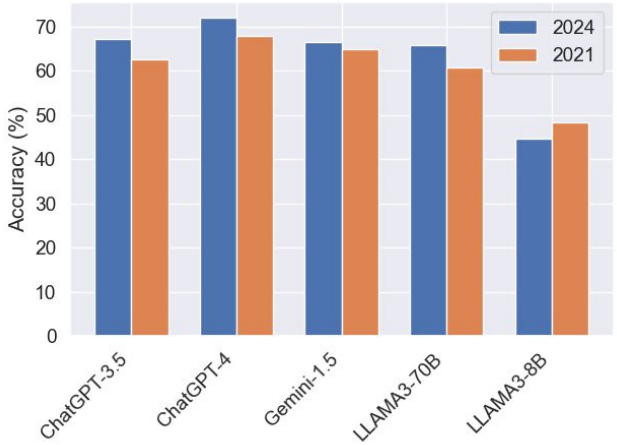
Correct Answer: T1059 (Command and Scripting Interpreter), T1573 (Encrypted Channel), T1546 (Event Triggered Execution)

Results Summary

Model	CTI-MCQ (Acc)	CTI-RCM (Acc)	CTI-VSP (MAD)	CTI-ATE (Macro-F1)	CTI-TAA (Acc)	
					Correct	Plausible
ChatGPT-4	**71.0**	**72.0**	1.31	**0.6388**	**52**	**86**
ChatGPT-3.5	54.1	67.2	1.57	0.3108	44	62
Gemini-1.5	65.4	66.6	**1.09**	0.4612	38	74
LLAMA3-70B	65.7	65.9	1.83	0.4720	52	80
LLAMA3-8B	61.3	44.7	1.91	0.1562	28	36

Performance of various LLMs on the CTIBench tasks. Acc refers to accuracy, MAD to mean absolute deviation (lower is better). For CTI-TAA, Correct means the LLM accurately identifies the threat actor or one of its aliases. Plausible refers to cases where the LLM provides a related or plausible threat actor when the report lacks sufficient detail to identify the exact actor. Bold values indicate the best-performing model.

Performance Before & After Knowledge Cutoff



Model	Before (F1)	After (F1)
ChatGPT-4	0.6542	0.6208
ChatGPT-3.5	0.3420	0.3333
Gemini-1.5	0.4360	0.5263
LLAMA3-70B	0.4934	0.4297
LLAMA3-8B	0.1813	0.1366

(Left): Performance on CTI-RCM and (Right) Performance on CTI-ATE before and after knowledge cutoff

Thank You!

Data: https://huggingface.co/datasets/AI4Sec/cti-bench
Code: https://github.com/xashru/cti-bench