Playing Defense: Benchmarking Cybersecurity Capabilities of Large Language Models

Adarsh Kyadige
Research Manager, Sophos AI

Co-authors:
Salma Taoufiq, Younghoo Lee, Tamas Voros, Konstantin Berlin
Agenda

- Introduction
  - Background
  - Contenders for our Benchmarks
- Our proposed Benchmarks
  - Natural Language Interface for Threat Hunting / Investigation
  - Incident Summarization
  - Artifact / Incident Evaluation
  - Benchmark Results and Discussion
- Takeaways
- Appendix
  - Detailed Results
  - Other considerations
    - Cost
    - Context size
Background
Large Language Models

Figure 1: The Transformer - model architecture.

http://jalammar.github.io/illustrated-transformer/
<table>
<thead>
<tr>
<th>T</th>
<th>Model</th>
<th>Average</th>
<th>ARC</th>
<th>HellaSwag</th>
<th>NMLU</th>
<th>TruthfulQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADC-ai-business/Marcoroni-700-v1</td>
<td>74.06</td>
<td>73.55</td>
<td>87.62</td>
<td>70.67</td>
<td>64.41</td>
</tr>
<tr>
<td></td>
<td>JCBU-NPU/FashionGPT-700-V1.1</td>
<td>74.05</td>
<td>71.76</td>
<td>88.2</td>
<td>70.99</td>
<td>65.26</td>
</tr>
<tr>
<td></td>
<td>adonlee/LLaMA_2_700_LoRA</td>
<td>73.9</td>
<td>72.7</td>
<td>87.55</td>
<td>70.84</td>
<td>64.52</td>
</tr>
<tr>
<td></td>
<td>uni-tianyan/Uni-TianYan</td>
<td>73.81</td>
<td>72.1</td>
<td>87.4</td>
<td>69.91</td>
<td>65.81</td>
</tr>
<tr>
<td></td>
<td>Riid/sheep-duck-llama-2</td>
<td>73.69</td>
<td>72.35</td>
<td>87.78</td>
<td>70.82</td>
<td>63.8</td>
</tr>
<tr>
<td></td>
<td>Riid/sheep-duck-llama-2</td>
<td>73.67</td>
<td>72.27</td>
<td>87.78</td>
<td>70.81</td>
<td>63.8</td>
</tr>
<tr>
<td></td>
<td>fangloveskari/ORCA_LLaMA_700_QLoRA</td>
<td>73.4</td>
<td>72.27</td>
<td>87.74</td>
<td>70.23</td>
<td>63.37</td>
</tr>
<tr>
<td></td>
<td>JCBU-NPU/FashionGPT-700-V1</td>
<td>73.26</td>
<td>71.08</td>
<td>87.32</td>
<td>70.7</td>
<td>63.92</td>
</tr>
<tr>
<td></td>
<td>oh-yeontaek/llama-2-700-LoRA-assemble-v2</td>
<td>73.22</td>
<td>71.84</td>
<td>86.89</td>
<td>69.37</td>
<td>64.79</td>
</tr>
<tr>
<td></td>
<td>hudecnxsystem/gpt2-70h</td>
<td>73.71</td>
<td>71.43</td>
<td>87.99</td>
<td>70.78</td>
<td>63.66</td>
</tr>
</tbody>
</table>

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
## The Main Contenders

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Size</th>
<th>Provider</th>
<th>Max. Context Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4</td>
<td>1.76T?</td>
<td>OpenAI</td>
<td>8k or 32k</td>
</tr>
<tr>
<td>GPT-3.5-Turbo</td>
<td>175B?</td>
<td></td>
<td>4k or 16k</td>
</tr>
<tr>
<td>Jurassic2-Ultra</td>
<td>?</td>
<td>Al21labs</td>
<td>8k</td>
</tr>
<tr>
<td>Jurassic2-Mid</td>
<td>?</td>
<td></td>
<td>8k</td>
</tr>
<tr>
<td>Claude-Instant</td>
<td>?</td>
<td>ANTHROPV\C</td>
<td>100k</td>
</tr>
<tr>
<td>Claude-v2</td>
<td>?</td>
<td></td>
<td>100k</td>
</tr>
<tr>
<td>Amazon-Titan-Large</td>
<td>45B</td>
<td>Amazon</td>
<td>4k</td>
</tr>
<tr>
<td>MPT-30B-Instruct</td>
<td>30B</td>
<td>mosaicML</td>
<td>8k</td>
</tr>
<tr>
<td>LLaMa2 (Chat-HF)</td>
<td>7B, 13B, 70B</td>
<td>Meta</td>
<td>4k</td>
</tr>
<tr>
<td>CodeLLaMa</td>
<td>7B, 13B, 34B</td>
<td></td>
<td>4k</td>
</tr>
</tbody>
</table>
Security Benchmarks for LLMs
The Three Benchmarks

Threat Hunting and Investigation assistance
- Input: Schema information and Natural language query
- Output: SQL query retrieving requested information

Incident Summarization
- Input: Alerts and User Activity data
- Output: Summary of the suspicious events, including extraction of important artifacts such as command lines, files and usernames

Artifact / Incident Evaluation
- Input: Alerts and User Activity data
- Output: Verdict on how malicious the input activity is, on a scale of 5 severity levels
Natural Language Interface
Natural Language Interface for Threat Hunting and Investigation

### Translate the following request into SQL

### Schema for alert_table table
<Table schema>

### Schema for process_table table
<Table schema>

### Schema for network_table table
<Table schema>

### Request:
- tell me a list of processes that were executed between 2021/10/19 and 2021/11/30
  SQL: `SELECT * FROM process_table WHERE timestamp BETWEEN '2021-10-19' AND '2021-11-30';`
- show me any low severity security alerts for the 23 days ago
  SQL: `SELECT * FROM alert_table WHERE severity='low' AND timestamp>=DATEADD('day', -23, CURRENT_TIMESTAMP());`
- show me the count of msword.exe processes that ran between Dec/01 and Dec/11
  SQL: `SELECT count(*) FROM process_table WHERE process='msword.exe' AND timestamp BETWEEN '2022-12-11' AND '2022-12-01';`
- Any ubuntu processes that was run by the user "admin" from host "db-server"
  SQL:
NLI Benchmark results

- Measures:
  - Query match
  - Result set match
  - Query Equivalence with GPT-4
- Top 4:
  - GPT-4
  - CodeLlama-34B-Instruct
  - Claude-v2
  - Claude-Instant
Give me a harder problem!

###
**Request:** tell me a list of processes that were executed between 2021/10/19 and 2021/11/30
SQL: `select * from process_table where timestamp between '2021-10-19' and '2021-11-30';`
###
**Request:** show me any low severity security alerts for the 23 days ago
SQL: `select * from alert_table where severity='low' and timestamp>=DATEADD('day', -23, CURRENT_TIMESTAMP());`
###
**Request:** show me the count of msword.exe processes that ran between Dec/01 and Dec/11
SQL: `select count(*) from process_table where process='msword.exe' and timestamp<= '2022-12-11' and timestamp >= '2022-12-01';`
###
**Request:** processes which were launched from the past 24 days
SQL: `select * from process_table where timestamp>DATEADD('day', -24, CURRENT_TIMESTAMP());`
###
**Request:** show me the top 3 most recently executed processes
SQL: `select * from process_table order by timestamp desc limit 3;`
###
**Request:** show me the top 3 most recently seen hosts
SQL: `select * from network_table order by timestamp desc limit 3;`
###
**Request:** show me the count of processes that ran on the 23rd of December
SQL: `select count(*) from process_table where timestamp='2022-12-12-23';`
###
**Request:** tell me the count of processes that ran on each day of the last week
SQL: `select count(*) from process_table where date(timestamp) in ( '2022-12-W01-1', '2022-12-W02-1', '2022-12-W03-1', '2022-12-W04-1', '2022-12-W05-1', '2022-12-W06-1', '2022-12-W07-1');`
Incident Summarization
Summarization

Summary:
<1-2 sentence description of what happened>

Observed MITRE Techniques:
<List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID>

Impacted Hosts:
<List of all hostname observed in the detections, provide corresponding IPs if available>

Active Users:
<List of all usernames observed in the detections. There could be multiple, list all of them>

Events:
<One sentence description for top three detection events. Start the list with 1.>
<Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that.>

Files:
<List the files found in the incident as follows:>
<TEMPLATE FOR FILES WITH DETAILS>

Command lines:
<List the command lines found in the detections as follows:>
<TEMPLATE FOR COMMAND LINES WITH DETAILS>
Summarization

Summary:
<1–2 sentence description of what happened>

Observed MITRE Techniques:
/List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID/>

Impacted Hosts:
/List of all hostname observed in the detections, provide corresponding IPs if available/>

Active Users:
/List of all usernames observed in the detections. There could be multiple, list all of them/>

Events:
/One sentence description for top three detection events. Start the list with 1. />
/Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that/>

Files:
/List the files found in the incident as follows:>
/<TEMPLATE FOR FILES WITH DETAILS/>

Command lines:
/List the command lines found in the detections as follows:>
/<TEMPLATE FOR COMMAND LINES WITH DETAILS/>
Summarization

Summary:
<1-2 sentence description of what happened>

Observed MITRE Techniques:
<List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID>

Impacted Hosts:
<List of all hostname observed in the detections, provide corresponding IPs if available>

Active Users:
<List of all usernames observed in the detections. There could be multiple, list all of them>

Events:
<One sentence description for top three detection events. Start the list with 1.>
<Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that.>

Files:
<List the files found in the incident as follows:>
<TEMPLATE FOR FILES WITH DETAILS>

Command lines:
<List the command lines found in the detections as follows:>
<TEMPLATE FOR COMMAND LINES WITH DETAILS>
Summary:
<1-2 sentence description of what happened>

Observed MITRE Techniques:
<List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID>

Impacted Hosts:
<List of all hostname observed in the detections, provide corresponding IPs if available>

Active Users:
<List of all usernames observed in the detections. There could be multiple, list all of them>

Events:
<One sentence description for top three detection events. Start the list with 1. >
<Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that.>

Files:
<List the files found in the incident as follows:>  
<TEMPLATE FOR FILES WITH DETAILS>

Command lines:
<List the command lines found in the detections as follows:>  
<TEMPLATE FOR COMMAND LINES WITH DETAILS>
Summarization

Summary:
<1-2 sentence description of what happened>

Observed MITRE Techniques:
<List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID>

Impacted Hosts:
<List of all hostname observed in the detections, provide corresponding IPs if available>

Active Users:
<List of all usernames observed in the detections. There could be multiple, list all of them>

Events:
<One sentence description for top three detection events. Start the list with 1. >
<Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that.>

Files:
<List the files found in the incident as follows:>
<TEMPLATE FOR FILES WITH DETAILS>

Command lines:
<List the command lines found in the detections as follows:>
<TEMPLATE FOR COMMAND LINES WITH DETAILS>
Summarization

Summary:  
<1-2 sentence description of what happened>

Observed MITRE Techniques:  
<List of all the observed mitre attack techniques found in the list of detections with technique and tactic ID>

Impacted Hosts:  
<List of all hostname observed in the detections, provide corresponding IPs if available>

Active Users:  
<List of all usernames observed in the detections. There could be multiple, list all of them>

Events:  
<One sentence description for top three detection events. Start the list with 1. >  
<Enumerate only up to ten artifacts under each report category, and summarize any remaining events beyond that >

Files:  
<List the files found in the incident as follows:>  
<TEMPLATE FOR FILES WITH DETAILS>

Command lines:  
<List the command lines found in the detections as follows:>  
<TEMPLATE FOR COMMAND LINES WITH DETAILS>
Incident Summarization Results
Incident Summarization Results
Incident Summary Comparison - GPT4

Summary:
The customer experienced multiple static exploit file detections, a brute force attempt, and an account lockout.

Observed MITRE Techniques:
- TA0002: Execution
- T1059: Command and Scripting Interpreter

Summary:
Multiple instances of suspicious activity were detected, including an attempt at a brute force attack on user accounts and suspicious files detected on the system, signaling a potential breach.

Observed MITRE Techniques:
- TA0002: Execution
- T1059: Command and Scripting Interpreter
Incident Summary Comparison – LLama-70B

Summary:
The customer experienced multiple static exploit file detections, a brute force attempt, and an account lockout.

Observed MITRE Techniques:
- TA0002: Execution
  - T1059: Command and Scripting Interpreter

Summary:
An attacker was detected attempting to brute-force attack a Windows account using invalid login credentials. Additionally, a suspicious file was detected on the system, which may indicate further malicious activity.

Observed MITRE Techniques:
* TA0002 - Execution
* T1059 - Command and Scripting Interpreter
Incident Summary Comparison – J2-Ultra

Summary:
The customer experienced multiple static exploit file detections, a brute force attempt, and an account lockout.

Observed MITRE Techniques:
- TA0002: Execution
  - T1059: Command and Scripting Interpreter

Summary:
A user account was locked out, a brute force attack was detected, and static exploit files were detected.

Observed MITRE Techniques:
1. Execution: Command and Scripting Interpreter (T1059)
2. Execution: Command and Scripting Interpreter (T1059)
3. Execution: Command and Scripting Interpreter (T1059)
Incident Summary Comparison – MPT-30B-Instruct

Summary:

The customer experienced multiple static exploit file detections, a brute force attempt, and an account lockout.

Observed MITRE Techniques:
- TA0002: Execution
  - T1059: Command and Scripting Interpreter

A user account was locked out from the [redacted] machine, with the source IP [redacted] shuffling through 20 or more different usernames, indicating a brute force attack. In addition, multiple exploit files were detected on the endpoint from [redacted] downloads, including: Exp/JPKGID-A, Exp/Datbi-A, Troj/SWF-CW, and two additional exploits."
Incident Summary Comparison – CodeLlama-34B-Instruct

Summary:
The customer experienced multiple static exploit file detections, a brute force attempt, and an account lockout.

Observed MITRE Techniques:
- TA0002: Execution
  - T1059: Command and Scripting Interpreter

<Regurgitates Detection Data>
Artifact / Incident Evaluation
Artifact/ Incident Evaluation

You are a helpful cybersecurity incident investigation expert that classifies incidents according to their severity level given a set of detections per incident.

Respond strictly with this JSON format: {"severity_pred": "xxx"} where xxx should only be either:
- Critical,
  <Criteria for a critical incident>
- High,
  <Criteria for a high severity incident>
- Medium,
  <Criteria for a medium severity incident>
- Low,
  <Criteria for a low severity incident>
- Informational
  <Criteria for an informational incident>

No other value is allowed.

Detections:
Artifact/Incident Evaluation Results
What we said

Respond strictly with this JSON format: {"severity_pred": "xxx"}
No other value is allowed.

### Do not respond with anything but the specified format

What the models did

```python
import pandas as pd
import numpy as np
import json

def classify_severity(detections):
    if len(detections) == 0:
        return "Informational"

    # Additional logic for severity classification

    if you can't determine the severity level of the incident, respond with {"severity_pred": "Undetermined"}
```

', '[2]._source.machine_data.columns.fileDescription': 'Microsoft @ Windows PowerShell',
'[2]._source.machine_data.columns.fileVersion': '10.0.19041.2913 (WinBuild.160101.0800)',
'[2]._source.machine_data.columns.name': 'powershell.exe', [2]._'
Putting it all together

Threat Hunting Assistance
- GPT-4
- CodeLlama-34B-Instruct
- Claude-v2

Incident Summarization
- GPT-4
- Claude-v2
- GPT-3.5-Turbo

Artifact Evaluation
- XGBoost w/ embeddings
- GPT-4
- Claude-v2
## Takeaways

<table>
<thead>
<tr>
<th><strong>Want a Threat Hunting assistant?</strong></th>
<th>You've got one right now!</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A little bit of prompt engineering will probably get you there</td>
</tr>
</tbody>
</table>

| **Want a model that summarizes unstructured data for quick review?** | Fine-tuning one of the better performers on this benchmark might yield a usable model. |

| **Want a model that can tell you if something bad is happening?** | There is no good model for this right now. A good intermediate solution is to use embeddings learned by the models. |
Appendix
Artifact/Incident evaluation – confusion matrices

XGBoost on GPT-3 embedding of incident-triggering IOC
Accuracy: 48.97%
Balanced accuracy: 26.51%

XGBoost model of initial human assessment
Accuracy: 24.36%
Balanced accuracy: 23.05%
Other important considerations
(Useful) Context size
# Natural Language Interface – Detailed results

<table>
<thead>
<tr>
<th>Model</th>
<th>Rank</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4</td>
<td>1</td>
<td>0.88</td>
</tr>
<tr>
<td>Claude-Instant</td>
<td>2</td>
<td>0.85</td>
</tr>
<tr>
<td>Claude-v2</td>
<td>2</td>
<td>0.85</td>
</tr>
<tr>
<td>CodeLlama-34B-Instruct</td>
<td>2</td>
<td>0.85</td>
</tr>
<tr>
<td>GPT-3.5-Turbo</td>
<td>5</td>
<td>0.84</td>
</tr>
<tr>
<td>Llama-2-70B-Chat</td>
<td>6</td>
<td>0.83</td>
</tr>
<tr>
<td>MPT-30B-Instruct</td>
<td>7</td>
<td>0.77</td>
</tr>
<tr>
<td>CodeLlama-7B-Instruct</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td>CodeLlama-13B-Instruct</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td>Jurassic-2-Ultra</td>
<td>10</td>
<td>0.72</td>
</tr>
<tr>
<td>Jurassic-2-Mid</td>
<td>11</td>
<td>0.57</td>
</tr>
<tr>
<td>Llama-2-13B-Chat</td>
<td>12</td>
<td>0.53</td>
</tr>
<tr>
<td>Amazon-Titan-Large</td>
<td>13</td>
<td>0.52</td>
</tr>
<tr>
<td>Llama-2-7B-Chat</td>
<td>14</td>
<td>0.39</td>
</tr>
</tbody>
</table>
### Summarization – Detailed results

<table>
<thead>
<tr>
<th>Model</th>
<th>Factual Accuracy</th>
<th>BERTScore (F1)</th>
<th>Cosine Similarity</th>
<th>METEOR score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LCS-based</td>
<td>Levenshtein-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>0.7646</td>
<td>0.8162</td>
<td>0.932</td>
<td>0.9696</td>
</tr>
<tr>
<td>Claude v2</td>
<td>0.6336</td>
<td>0.7235</td>
<td>0.8514</td>
<td>0.9238</td>
</tr>
<tr>
<td>Claude Instant v1</td>
<td>0.5503</td>
<td>0.6488</td>
<td>0.8404</td>
<td>0.9126</td>
</tr>
<tr>
<td>J2-Ultra</td>
<td>0.527</td>
<td>0.6309</td>
<td>0.8359</td>
<td>0.9133</td>
</tr>
<tr>
<td>J2-Mid</td>
<td>0.2862</td>
<td>0.3707</td>
<td>0.8088</td>
<td>0.8774</td>
</tr>
<tr>
<td>LLaMA2 7B Chat HF</td>
<td>0.2229</td>
<td>0.3974</td>
<td>0.7782</td>
<td>0.8485</td>
</tr>
<tr>
<td>LLaMA2 13B Chat HF</td>
<td>0.3786</td>
<td>0.5602</td>
<td>0.8214</td>
<td>0.8705</td>
</tr>
<tr>
<td>LLaMA2 70B Chat HF</td>
<td>0.4739</td>
<td>0.6547</td>
<td>0.826</td>
<td>0.8847</td>
</tr>
<tr>
<td>CodeLLaMA 7B Instruct</td>
<td>0.1189</td>
<td>0.2194</td>
<td>0.7652</td>
<td>0.8083</td>
</tr>
<tr>
<td>CodeLLaMA 13B Instruct</td>
<td>0.3128</td>
<td>0.4962</td>
<td>0.7988</td>
<td>0.8488</td>
</tr>
<tr>
<td>CodeLLaMA 34B Instruct</td>
<td>0.252</td>
<td>0.3746</td>
<td>0.7917</td>
<td>0.8245</td>
</tr>
<tr>
<td>MPT30B Instruct</td>
<td>0.2594</td>
<td>0.3783</td>
<td>0.8178</td>
<td>0.8702</td>
</tr>
<tr>
<td>Amazon Titan Large</td>
<td>0.4642</td>
<td>0.5904</td>
<td>0.8418</td>
<td>0.9045</td>
</tr>
<tr>
<td>GPT 3.5 Turbo</td>
<td>0.5586</td>
<td>0.6808</td>
<td>0.8624</td>
<td>0.9323</td>
</tr>
<tr>
<td>Model</td>
<td>Accuracy</td>
<td>Balanced Accuracy</td>
<td>Per-class F1 score (Critical to Informational)</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------</td>
<td>------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>0.04</td>
<td>0.2537</td>
<td>[0.0245, 0.0478, 0.108, 0, 0]</td>
<td></td>
</tr>
<tr>
<td>GPT-3.5-Turbo</td>
<td>0.26</td>
<td>0.2336</td>
<td>[0.051, 0.873, 0.373, 0.270, 0.0076]</td>
<td></td>
</tr>
<tr>
<td>Claude-v2</td>
<td>0.2</td>
<td>0.2614</td>
<td>[0.035, 0.104, 0.33, 0.202, 0]</td>
<td></td>
</tr>
<tr>
<td>Claude-Instant-v1</td>
<td>0.07</td>
<td>0.2094</td>
<td>[0.013, 0.082, 0.218, 0.004, 0]</td>
<td></td>
</tr>
<tr>
<td>J2-Ultra</td>
<td>0.49</td>
<td>0.2083</td>
<td>[0.043, 0, 0, <strong>0.66</strong>, 0.038]</td>
<td></td>
</tr>
<tr>
<td>J2-Mid</td>
<td>0.19</td>
<td>0.1786</td>
<td>[0, 0, 0.307, 0, 0.221]</td>
<td></td>
</tr>
<tr>
<td>Amazon-Titan-Large</td>
<td>0.007</td>
<td>0.2001</td>
<td>[0.012, 0, 0, 0.002, 0]</td>
<td></td>
</tr>
<tr>
<td>Llama-2-7B-Chat</td>
<td>0.001</td>
<td>0.063</td>
<td>[0, 0, 0.267, 0.118, 0]</td>
<td></td>
</tr>
<tr>
<td>Llama-2-13B-Chat</td>
<td>0.001</td>
<td>0.13</td>
<td>[0, 0, 0, 0.67, 0]</td>
<td></td>
</tr>
<tr>
<td>Llama-2-70B-Chat</td>
<td>0.008</td>
<td>0.05</td>
<td>[0, 0.048, 0.112, 0.116, 0.283]</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-7B-Chat</td>
<td>0.02</td>
<td>0.129</td>
<td>[0, 0.078, 0.118, 0.321, 0.057]</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-13B-Chat</td>
<td>0.003</td>
<td>0.128</td>
<td>[0, 0.2, 0.23, 0.098, 0.077]</td>
<td></td>
</tr>
<tr>
<td>CodeLlama-34B-Chat</td>
<td>0.016</td>
<td>0.046</td>
<td>[0, 0, 0.075, 0.28, 0.28]</td>
<td></td>
</tr>
</tbody>
</table>
## Artifact Evaluation – Detailed Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Experiment</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Per-class F1 score (Critical to Informational)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4</td>
<td>3-Shot with distinct severities (case-creating IOC)</td>
<td>0.14</td>
<td><strong>0.2827</strong></td>
<td>[0.035, 0.070, 0.312, 0.069, 0]</td>
</tr>
<tr>
<td>GPT-3.5-Turbo</td>
<td>3-Shot with distinct severities (case-creating IOC)</td>
<td>0.21</td>
<td>0.2403</td>
<td>[0.04, 0.08, 0.32, 0.258, 0.004]</td>
</tr>
<tr>
<td>Claude-v2</td>
<td>3-Shot with distinct severities (case-creating IOC)</td>
<td>0.28</td>
<td>0.249</td>
<td>[0.038, <strong>0.95</strong>, 0.332, 0.38, 0.07]</td>
</tr>
<tr>
<td>XGBoost</td>
<td>GPT-3 embeddings of case-creating IOC</td>
<td><strong>0.49</strong></td>
<td>0.2651</td>
<td>[0.069, 0.014, 0.40, 0.614, 0.23]</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Amazon Embedding G1 embeddings of case-creating IOC</td>
<td>0.46</td>
<td>0.2646</td>
<td>[0.066, 0.0625, 0.452, 0.56, 0.16]</td>
</tr>
<tr>
<td>XGBoost</td>
<td>BGE-Large embeddings of case-creating IOC</td>
<td>0.46</td>
<td>0.2518</td>
<td>[0.064, 0.022, 0.407, 0.603, 0.126]</td>
</tr>
<tr>
<td>XGBoost</td>
<td>GPT-3 embeddings of multiple IOCs</td>
<td>0.48</td>
<td>0.2614</td>
<td>[<strong>0.077</strong>, 0.058, 0.367, 0.624, 0.205]</td>
</tr>
<tr>
<td>XGBoost</td>
<td>2 features: Incident-creating IOC’s severity and its type</td>
<td>0.2436</td>
<td>0.2305</td>
<td>[0.003, 0.0914, 0.3495, 0.23, 0.36]</td>
</tr>
</tbody>
</table>