# 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

SOPHOS

# Sophos AI

# 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

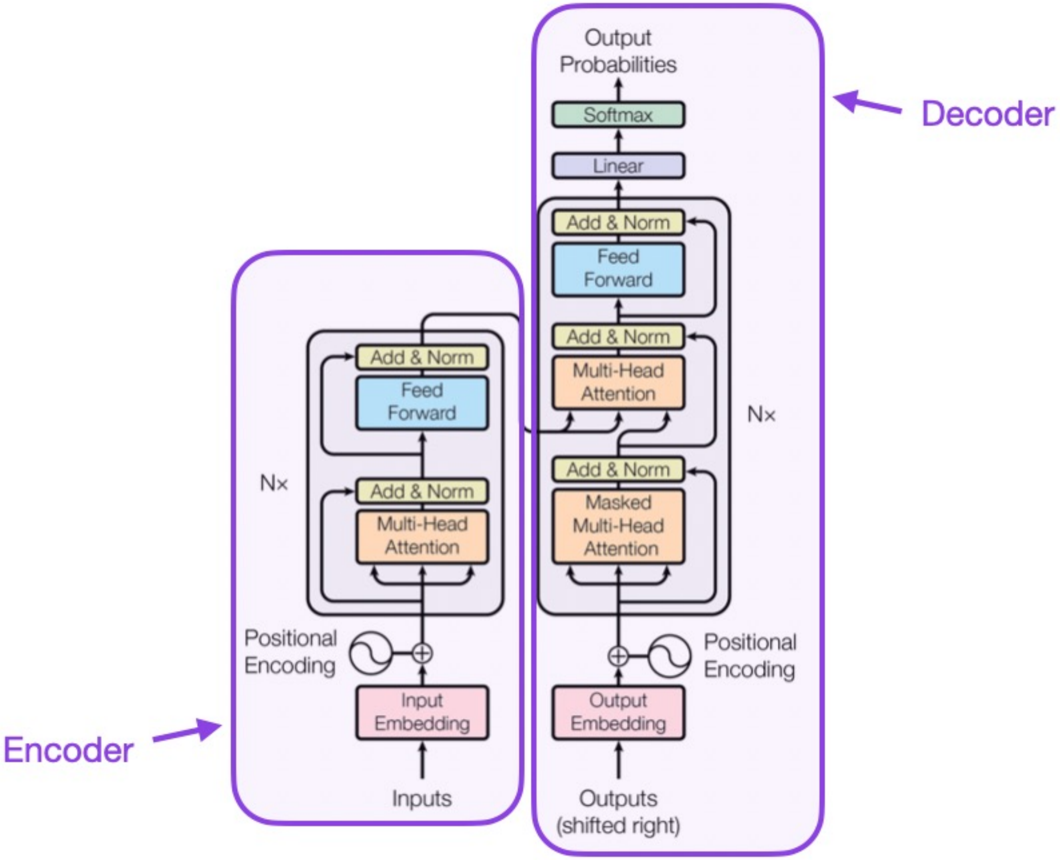
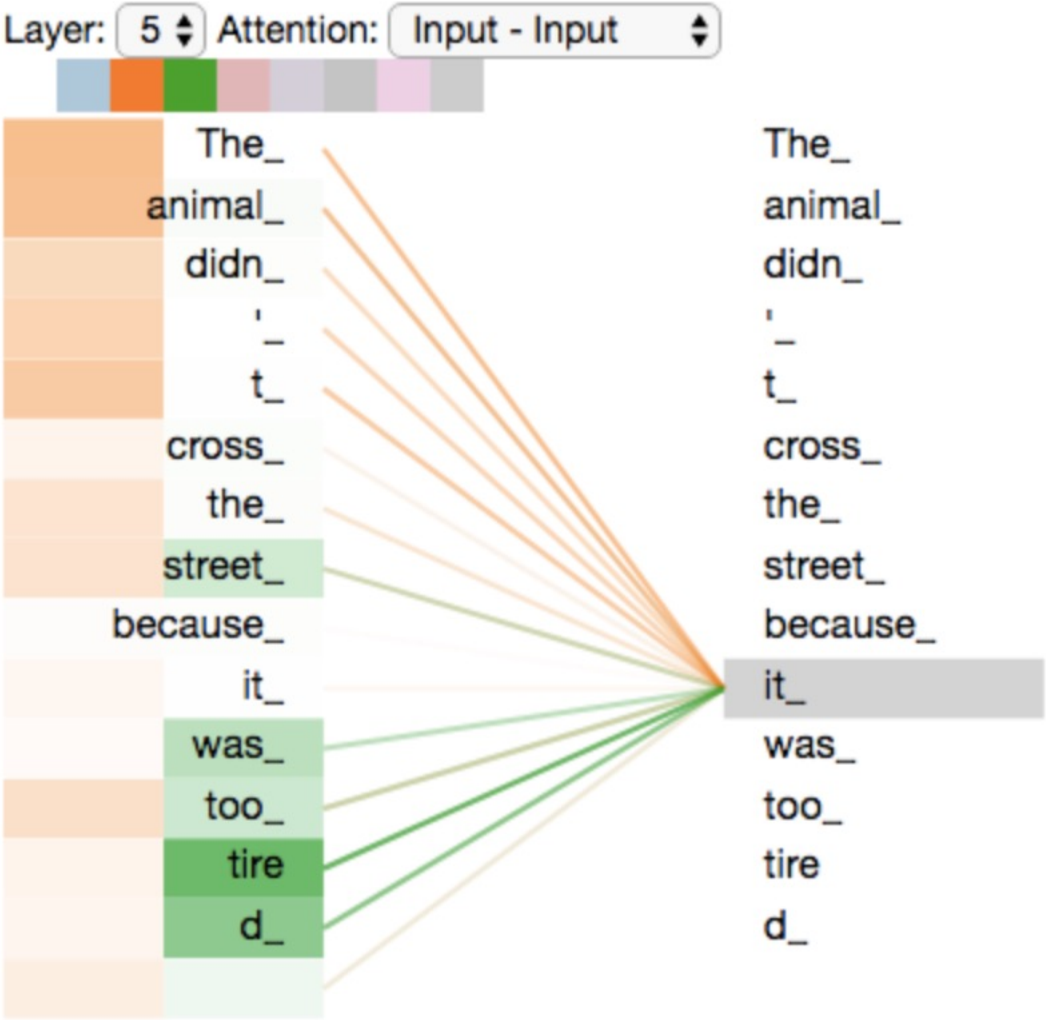
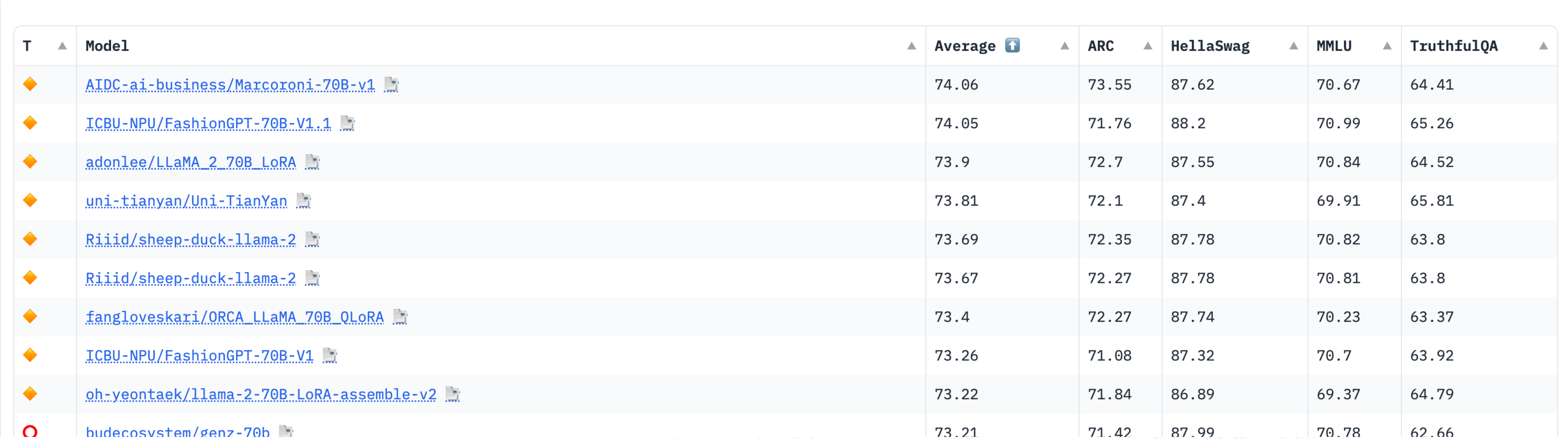


Figure 1: The Transformer - model architecture.



http://jalammar.github.io/illustrated-transformer/
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

# LLM Leaderboard



| T | Model | Average ⬆️ | ARC | HellaSwag | MMLU | TruthfulQA |
|---|---|---|---|---|---|---|
| 🔶 | AIDC-ai-business/Marcoroni-70B-v1 | 74.06 | 73.55 | 87.62 | 70.67 | 64.41 |
| 🔶 | ICBU-NPU/FashionGPT-70B-V1.1 | 74.05 | 71.76 | 88.2 | 70.99 | 65.26 |
| 🔶 | adonlee/LLaMA_2_70B_LoRA | 73.9 | 72.7 | 87.55 | 70.84 | 64.52 |
| 🔶 | uni-tianyan/Uni-TianYan | 73.81 | 72.1 | 87.4 | 69.91 | 65.81 |
| 🔶 | Riiid/sheep-duck-llama-2 | 73.69 | 72.35 | 87.78 | 70.82 | 63.8 |
| 🔶 | Riiid/sheep-duck-llama-2 | 73.67 | 72.27 | 87.78 | 70.81 | 63.8 |
| 🔶 | fangloveskari/ORCA_LLaMA_70B_QLoRA | 73.4 | 72.27 | 87.74 | 70.23 | 63.37 |
| 🔶 | ICBU-NPU/FashionGPT-70B-V1 | 73.26 | 71.08 | 87.32 | 70.7 | 63.92 |
| 🔶 | oh-yeontaek/llama-2-70B-LoRA-assemble-v2 | 73.22 | 71.84 | 86.89 | 69.37 | 64.79 |
| ⭕ | budecosystem/genz-70b | 73.21 | 71.42 | 87.99 | 70.78 | 62.66 |

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

# The Main Contenders

| Model Name | Size | Provider | Max. Context Window |
|---|---|---|---|
| **GPT-4** | 1.76T? | OpenAI | 8k or 32k |
| **GPT-3.5-Turbo** | 175B? | | 4k or 16k |
| **Jurassic2-Ultra** | ? | AI21 labs | 8k |
| **Jurassic2-Mid** | ? | | 8k |
| **Claude-Instant** | ? | ANTHROP\C | 100k |
| **Claude-v2** | ? | | 100k |
| **Amazon-Titan-Large** | 45B | amazon | 4k |
| **MPT-30B-Instruct** | 30B | mosaicML | 8k |
| **LlaMa2 (Chat-HF)** | 7B, 13B, 70B | Meta | 4k |
| **CodeLlaMa** | 7B, 13B, 34B | | 4k |

# 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

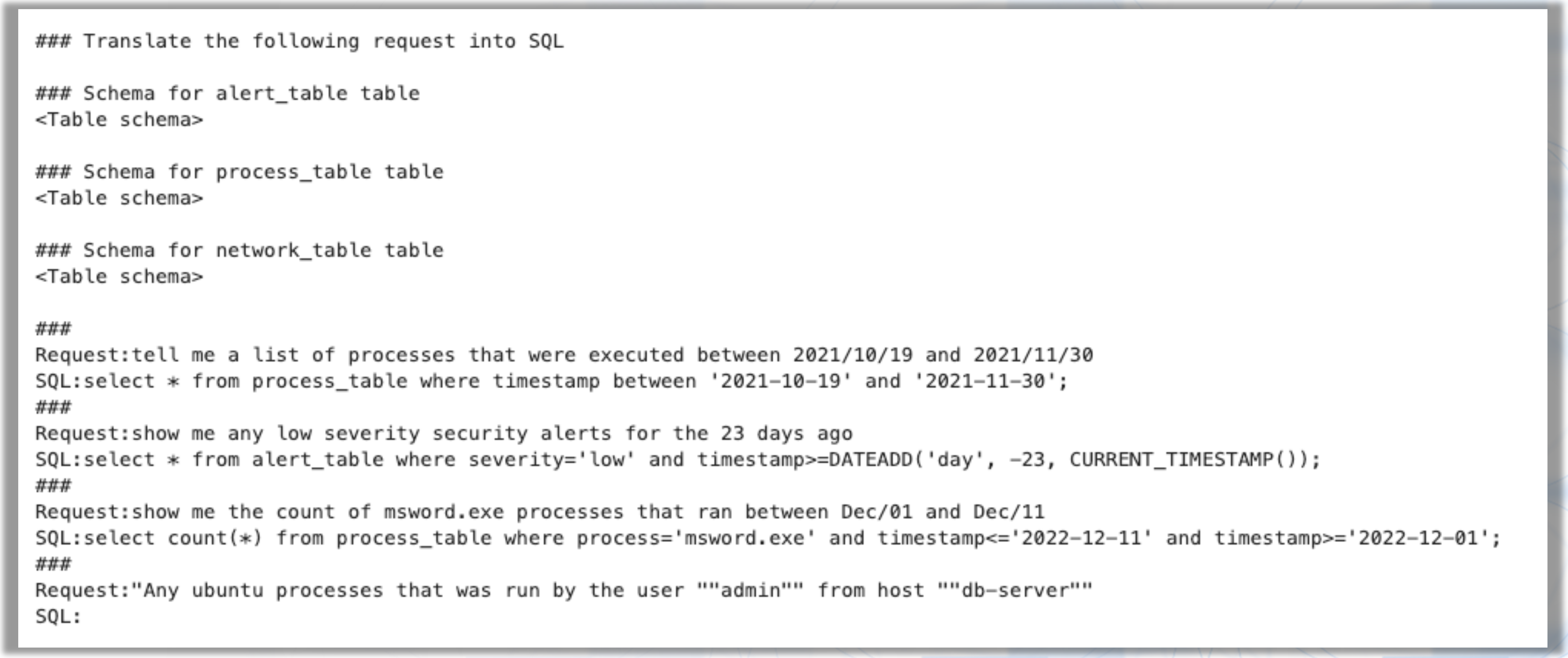
- Input: Alerts and User Activity data
- Output: Summary of the suspicious events, including extraction of important artifacts such as command lines, files and usernames

## 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

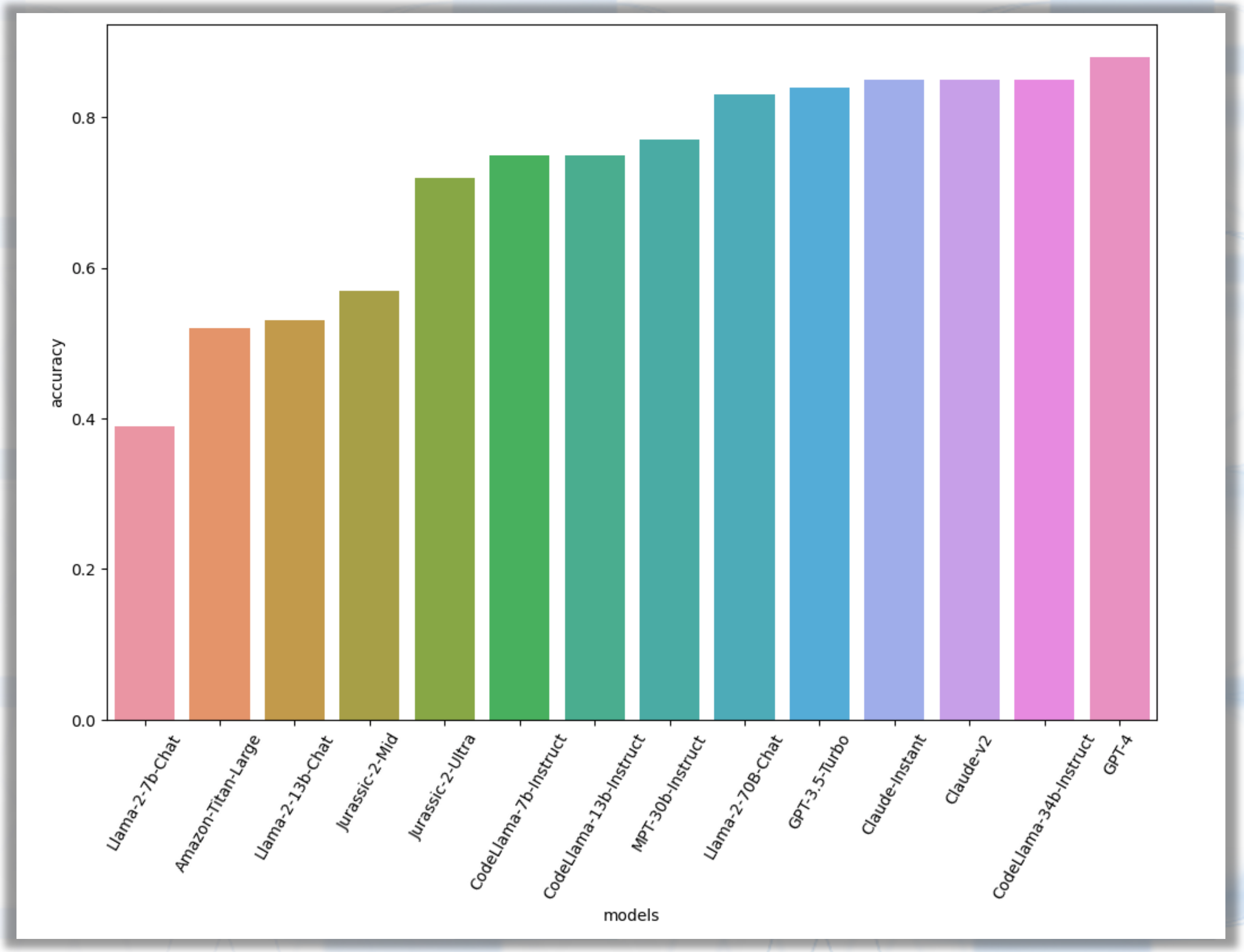
# Natural Language Interface for Threat Hunting and Investigation



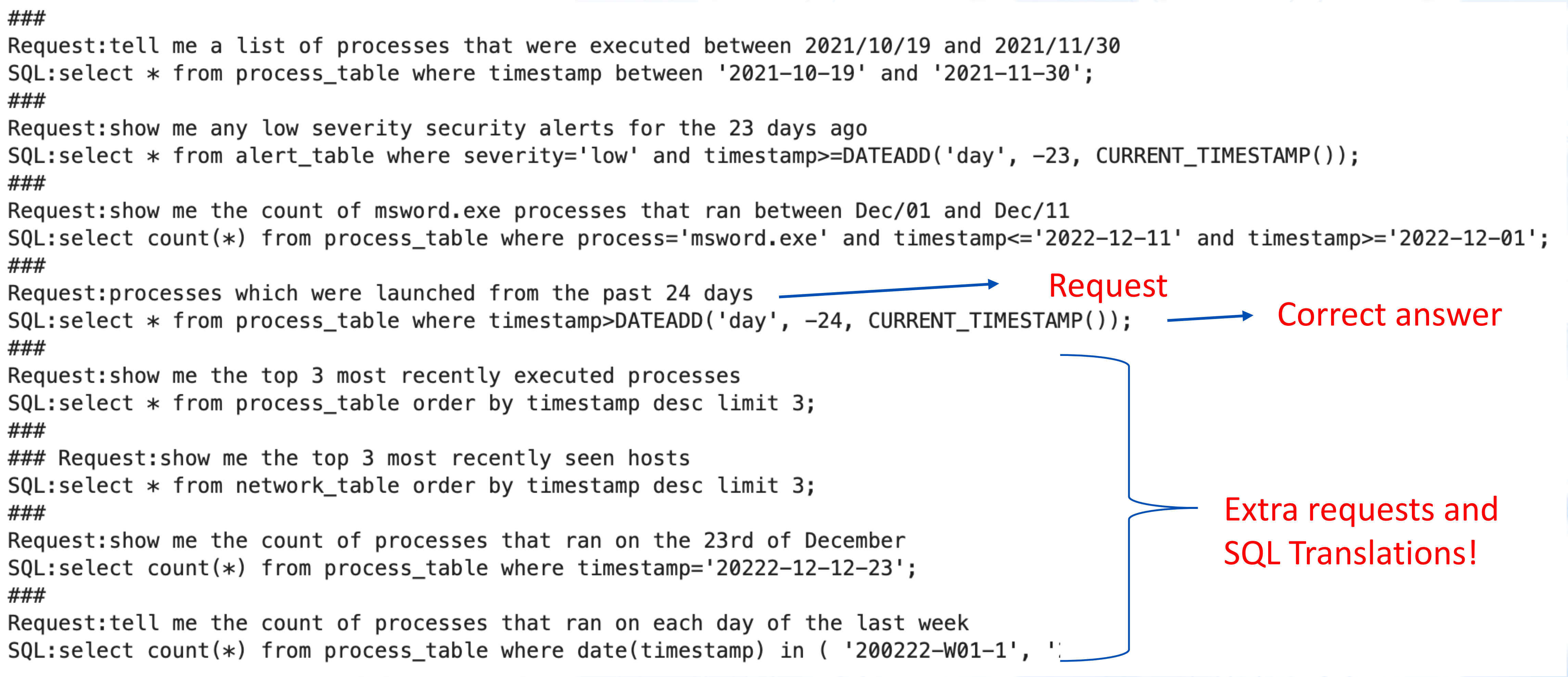
```
### Translate the following request into SQL

### 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';
###
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:"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='20222-12-12-23';
###
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 ( '200222-W01-1', '
```

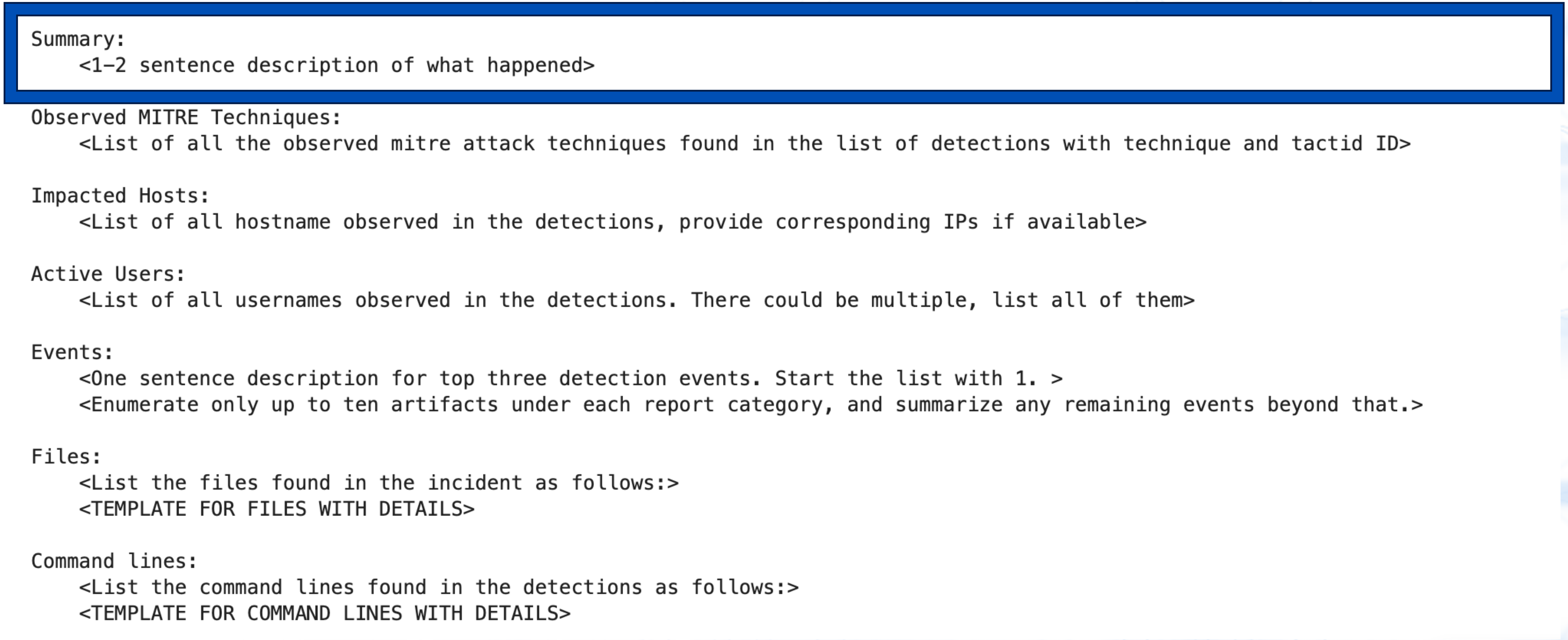
Request

Correct answer

Extra requests and
SQL Translations!

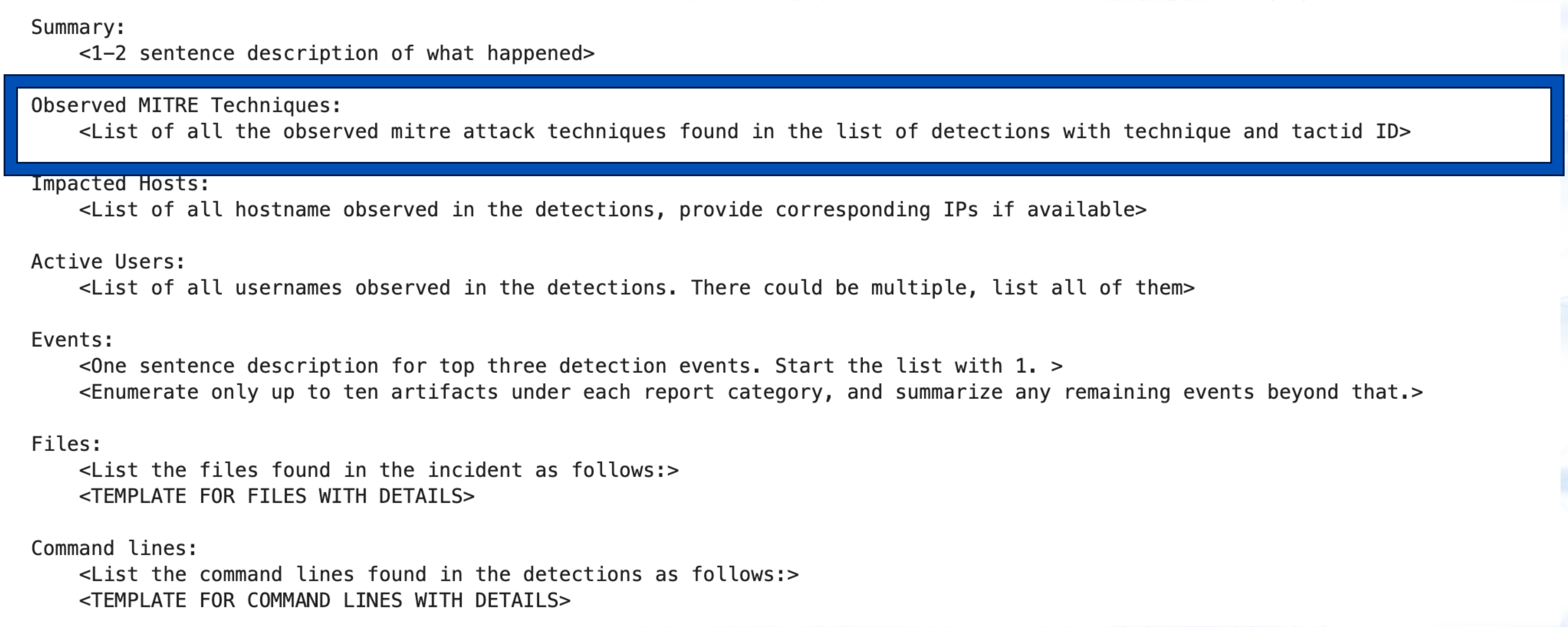
# Incident Summarization

# Summarization



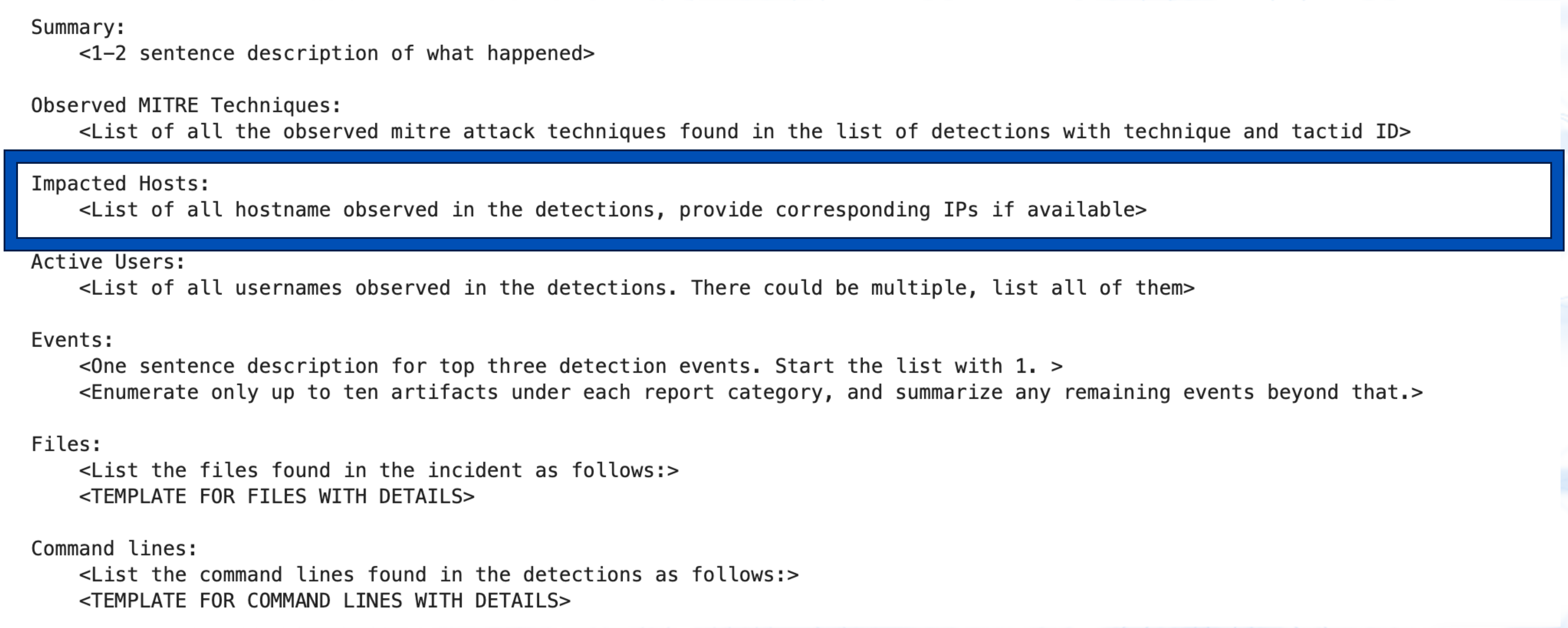
```
Summary:
    <1-2 sentence description of what happened>

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



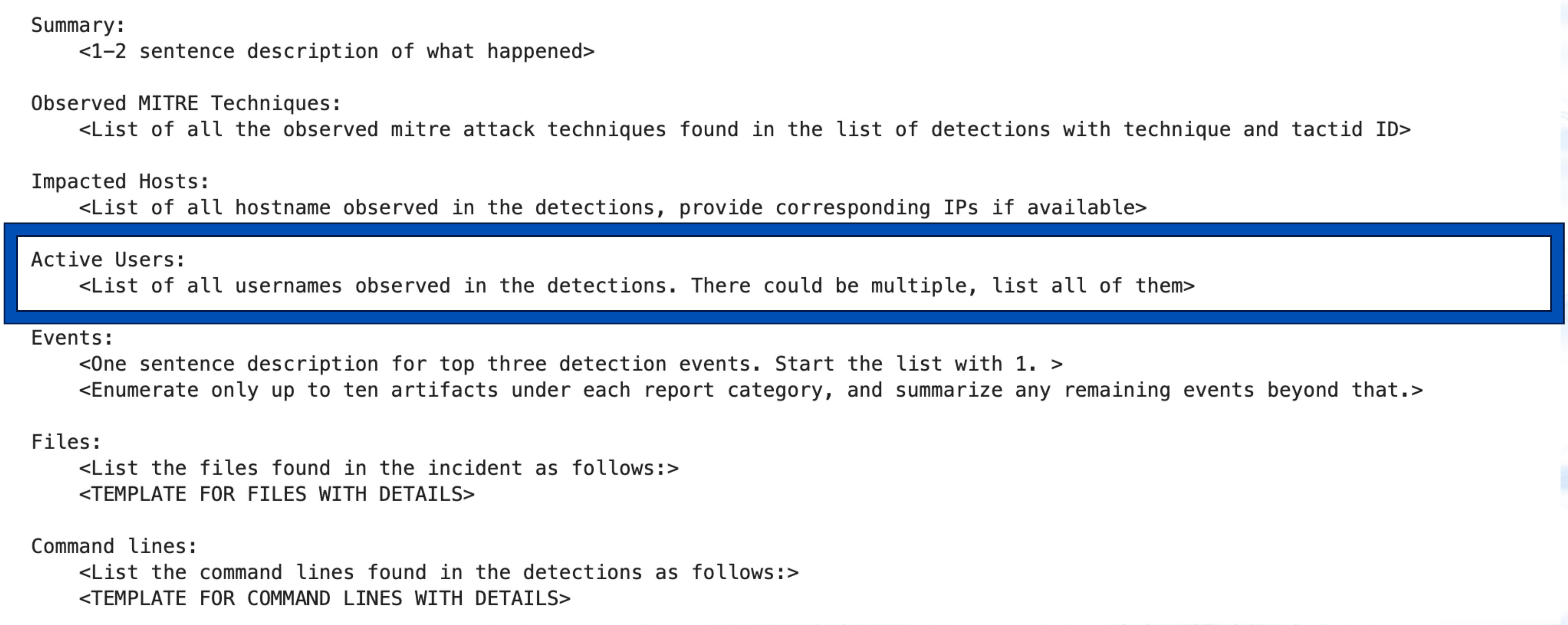
```
Summary:
    <1-2 sentence description of what happened>

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



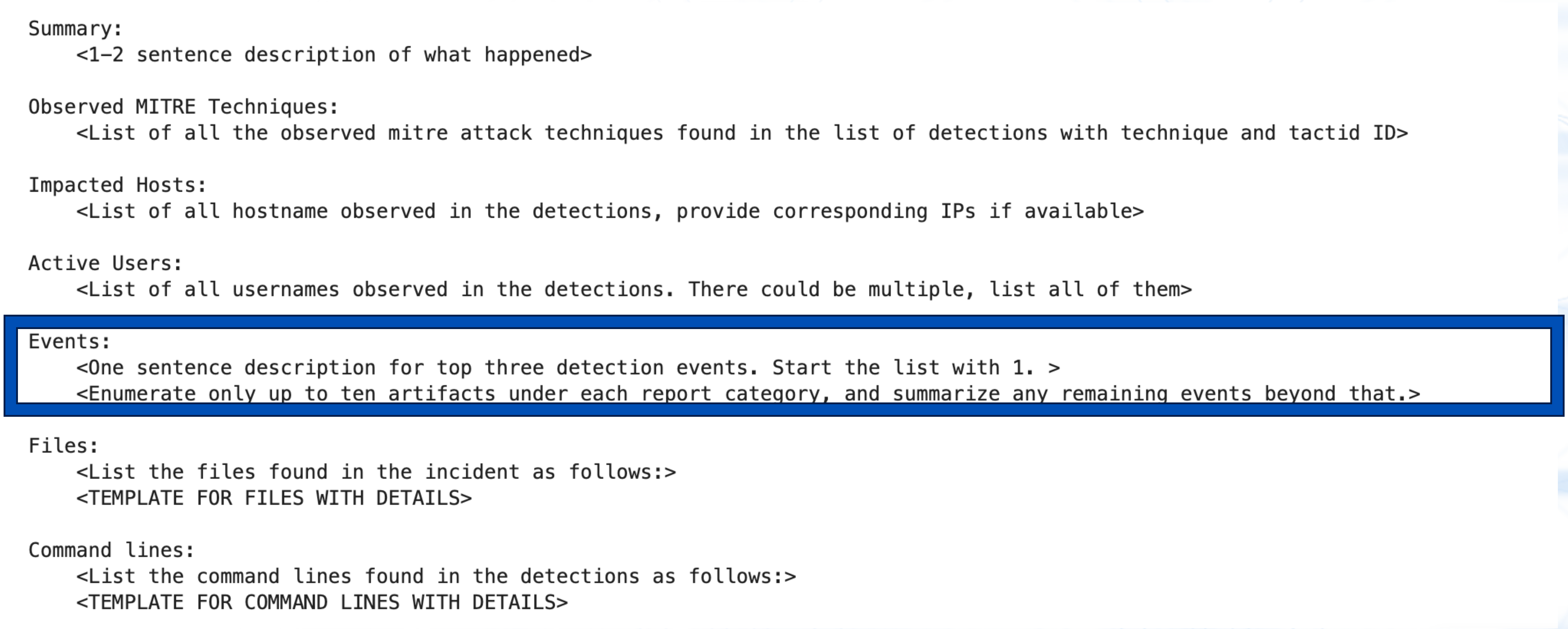
```
Summary:
    <1-2 sentence description of what happened>

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



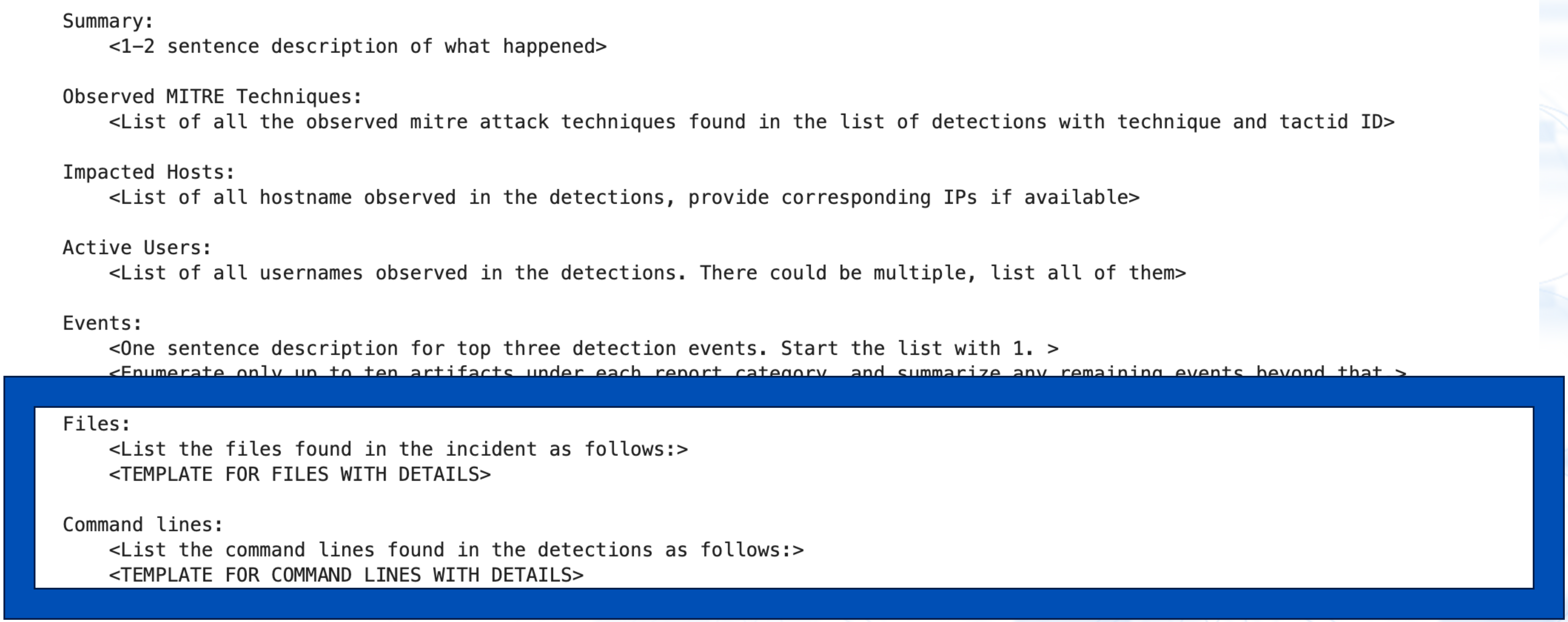
```
Summary:
    <1-2 sentence description of what happened>

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



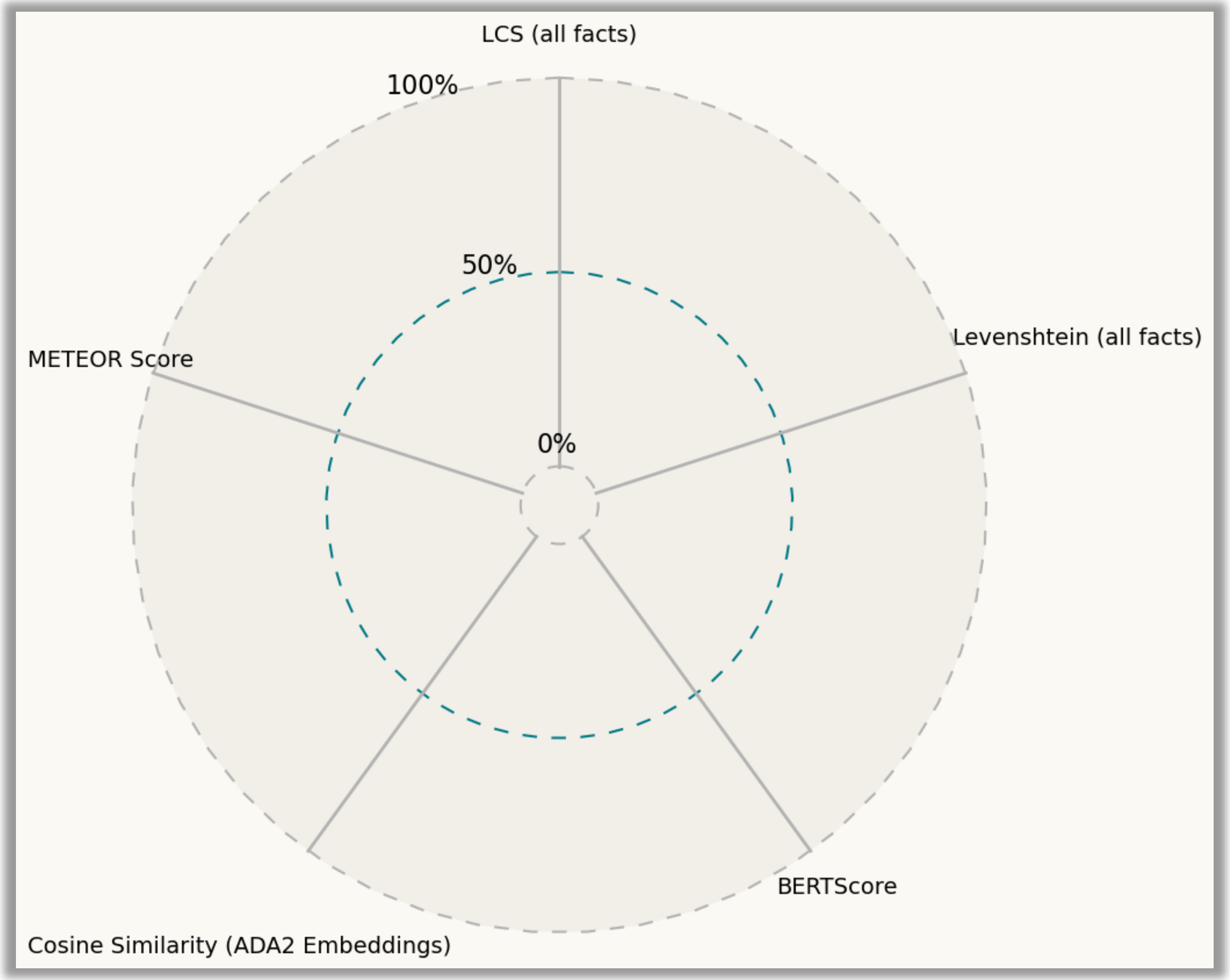
```
Summary:
    <1-2 sentence description of what happened>

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

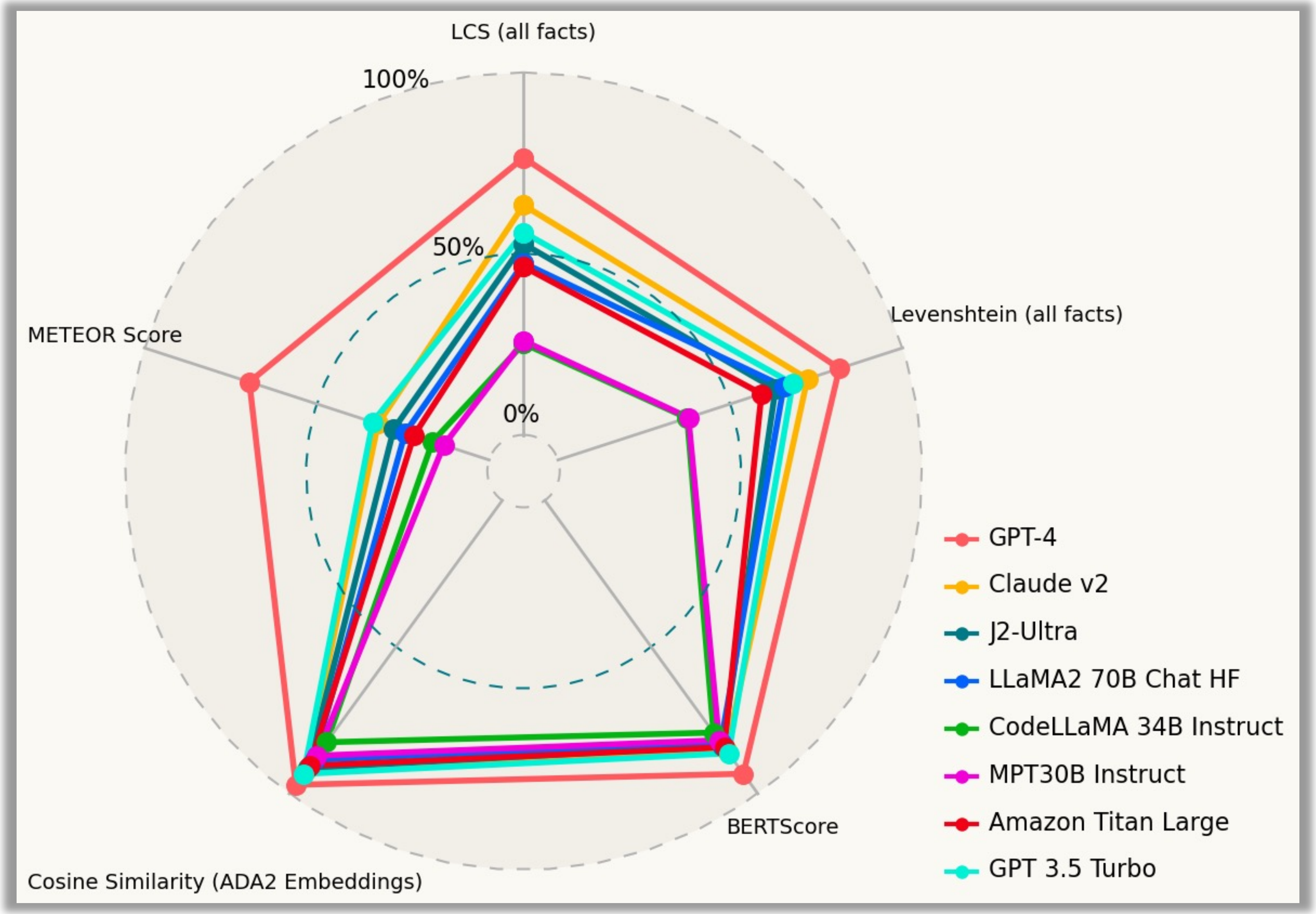


```
Summary:
    <1-2 sentence description of what happened>

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

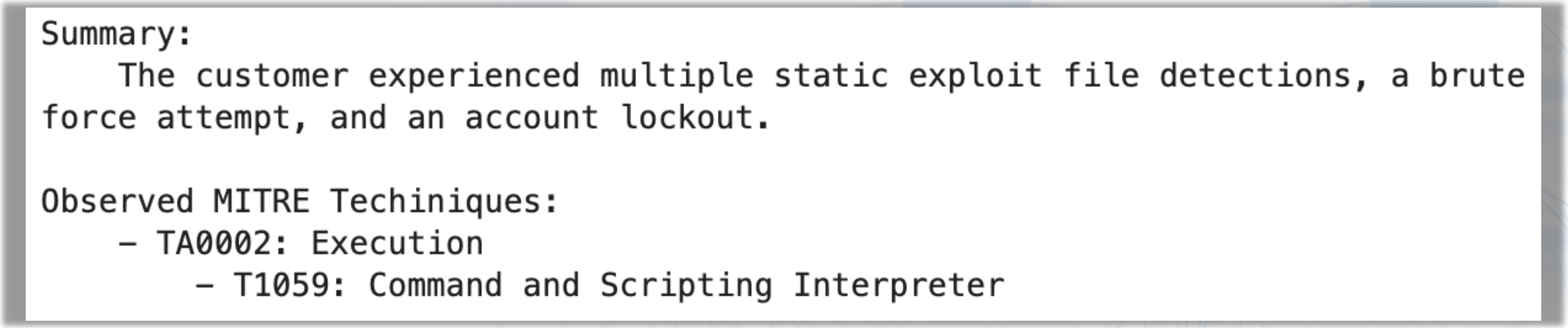
# Incident Summarization Results



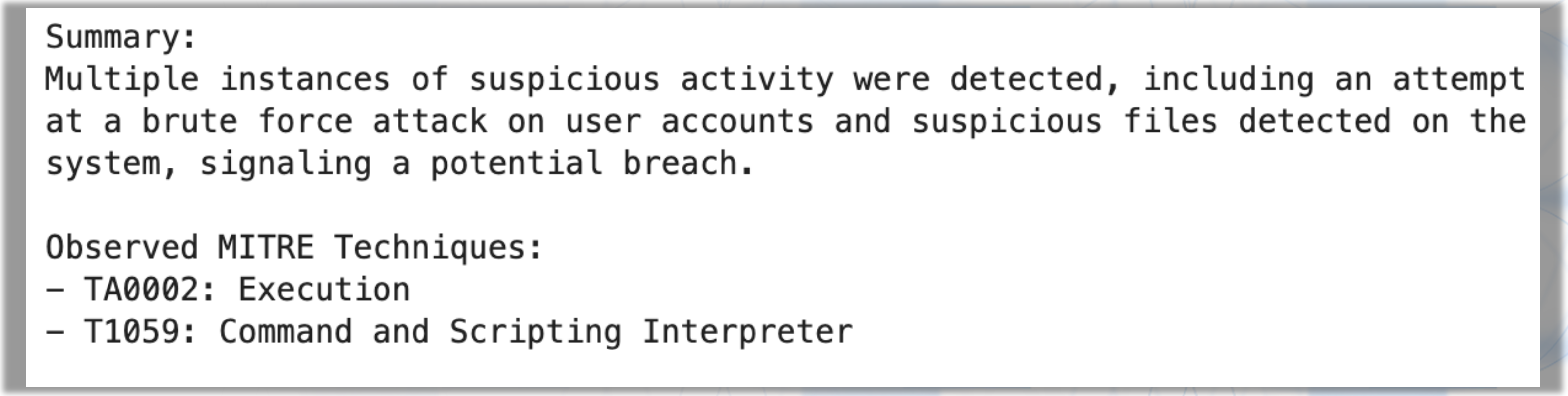
# Incident Summarization Results



# Incident Summary Comparison - GPT4

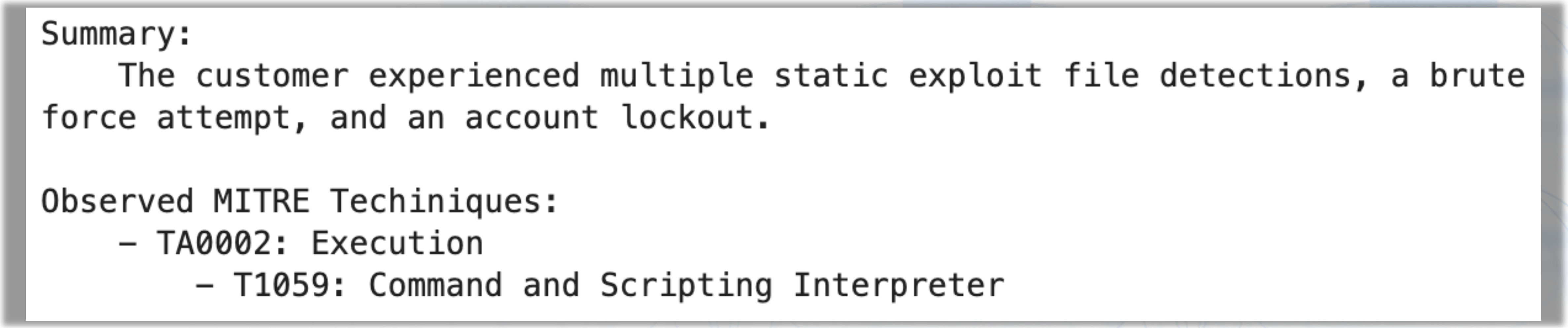


```
Summary:
    The customer experienced multiple static exploit file detections, a brute
force attempt, and an account lockout.

Observed MITRE Techiniques:
    - TA0002: Execution
        - T1059: Command and Scripting Interpreter
```

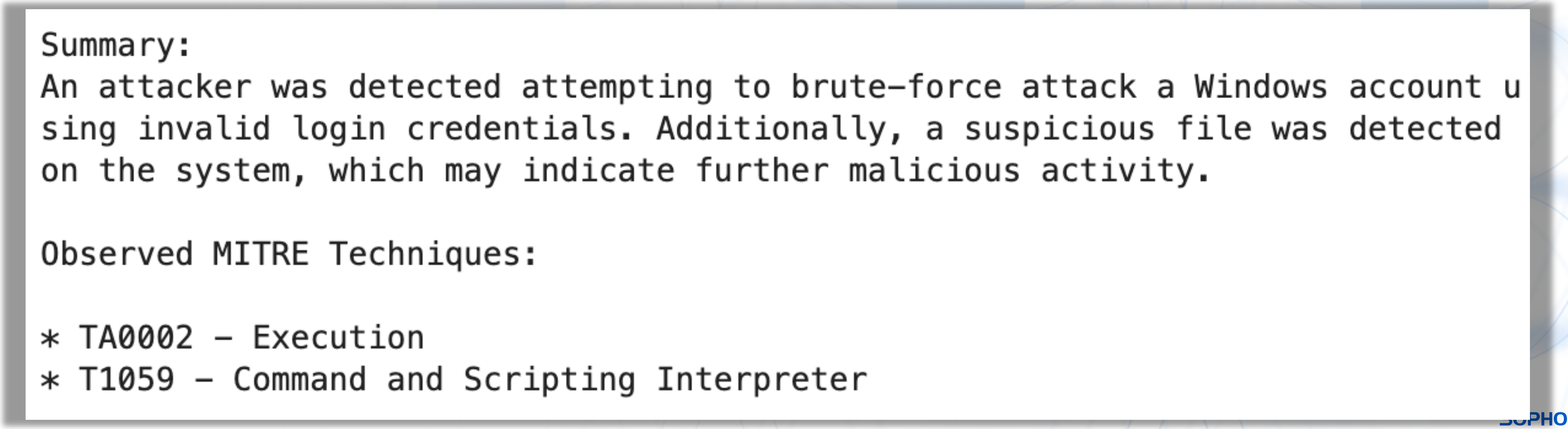


```
Summary:
Multiple instances of suspicious activity were detected, including an attempt
at a brute force attack on user accounts and suspicious files detected on the
system, signaling a potential breach.

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

# Incident Summary Comparison – LLama-70B

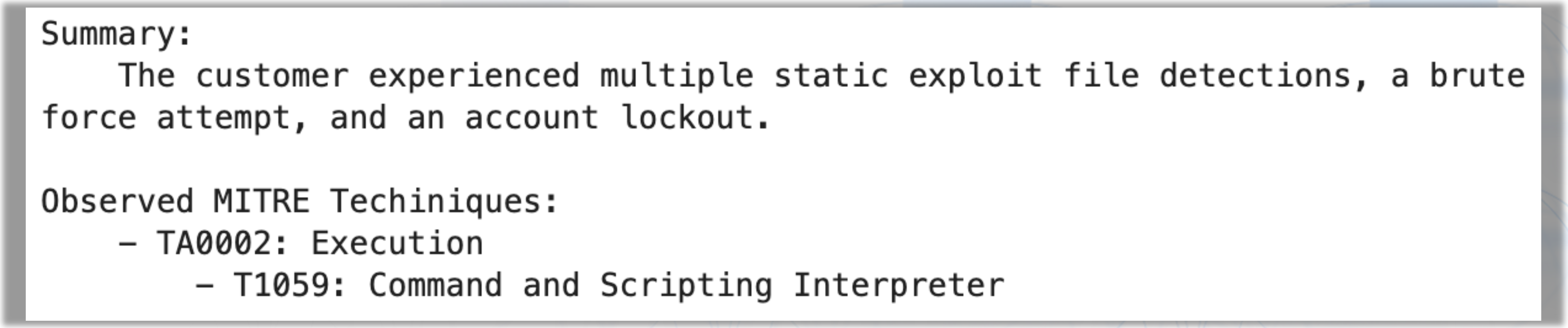


```
Summary:
    The customer experienced multiple static exploit file detections, a brute
force attempt, and an account lockout.

Observed MITRE Techiniques:
    - TA0002: Execution
        - T1059: Command and Scripting Interpreter
```

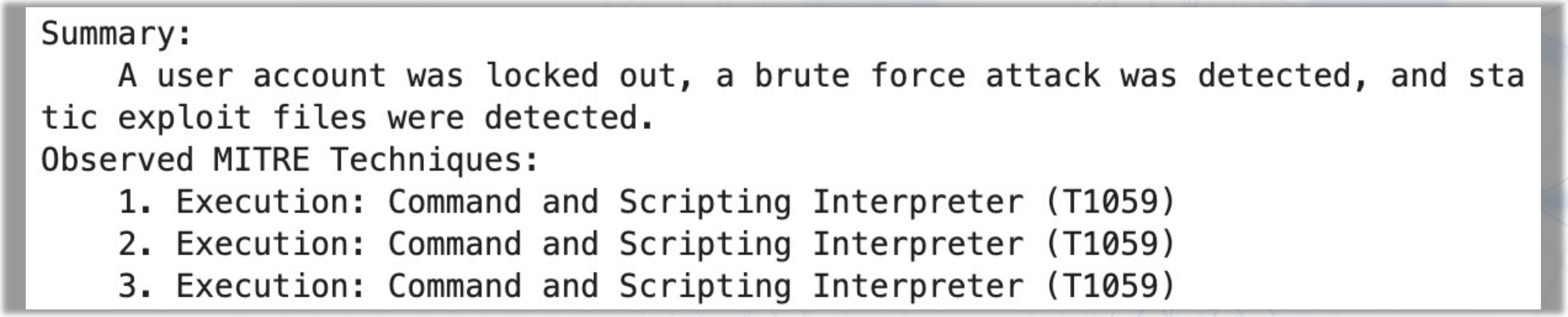


```
Summary:
An attacker was detected attempting to brute-force attack a Windows account u
sing invalid login credentials. Additionally, a suspicious file was detected
on the system, which may indicate further malicious activity.

Observed MITRE Techniques:

* TA0002 - Execution
* T1059 - Command and Scripting Interpreter
```

# Incident Summary Comparison – J2-Ultra

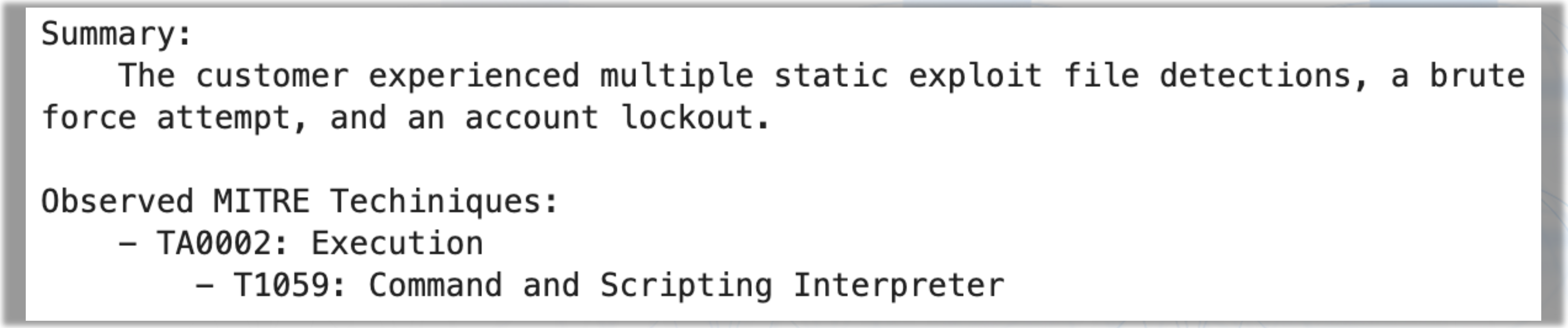


```
Summary:
    The customer experienced multiple static exploit file detections, a brute
force attempt, and an account lockout.

Observed MITRE Techiniques:
    - TA0002: Execution
        - T1059: Command and Scripting Interpreter
```

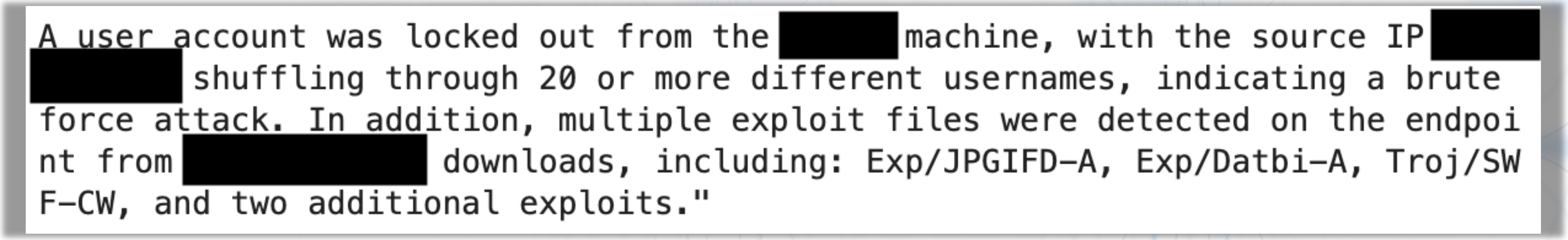


```
Summary:
    A user account was locked out, a brute force attack was detected, and sta
tic exploit files were detected.
Observed MITRE Techniques:
    1. Execution: Command and Scripting Interpreter (T1059)
    2. Execution: Command and Scripting Interpreter (T1059)
    3. Execution: Command and Scripting Interpreter (T1059)
```

# Incident Summary Comparison – MPT-30B-Instruct

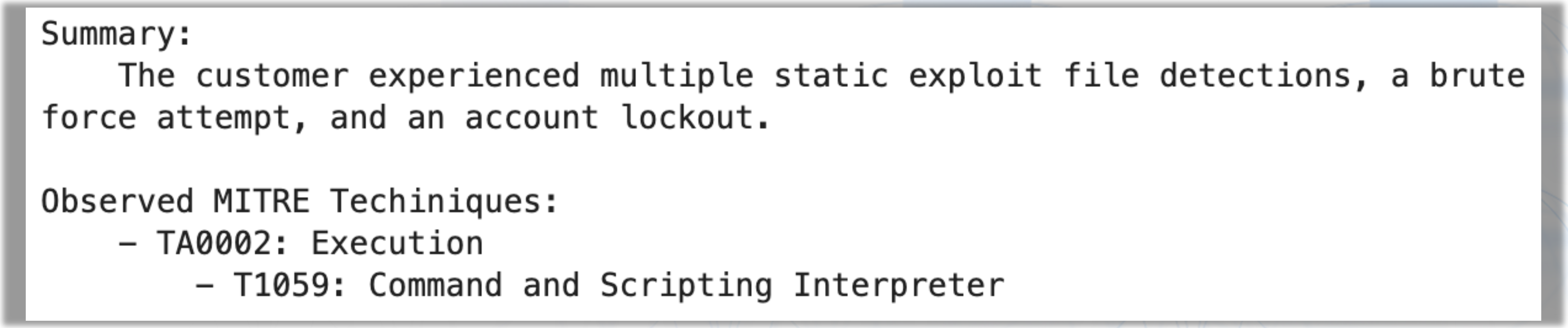


```
Summary:
    The customer experienced multiple static exploit file detections, a brute
force attempt, and an account lockout.

Observed MITRE Techiniques:
    - 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 endpoi
nt from [redacted] downloads, including: Exp/JPGIFD-A, Exp/Datbi-A, Troj/SW
F-CW, and two additional exploits."
```

# Incident Summary Comparison – CodeLlama-34B-Instruct

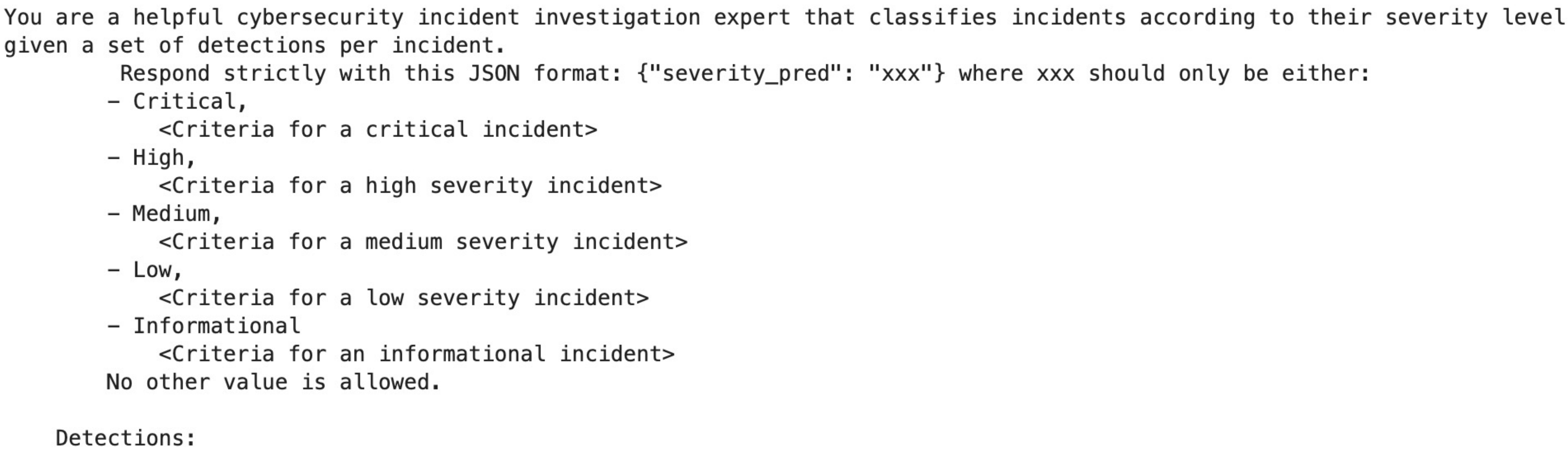


```
Summary:
    The customer experienced multiple static exploit file detections, a brute
force attempt, and an account lockout.

Observed MITRE Techiniques:
    - TA0002: Execution
        - T1059: Command and Scripting Interpreter
```

<Regurgitates Detection Data>

# Artifact / Incident Evaluation

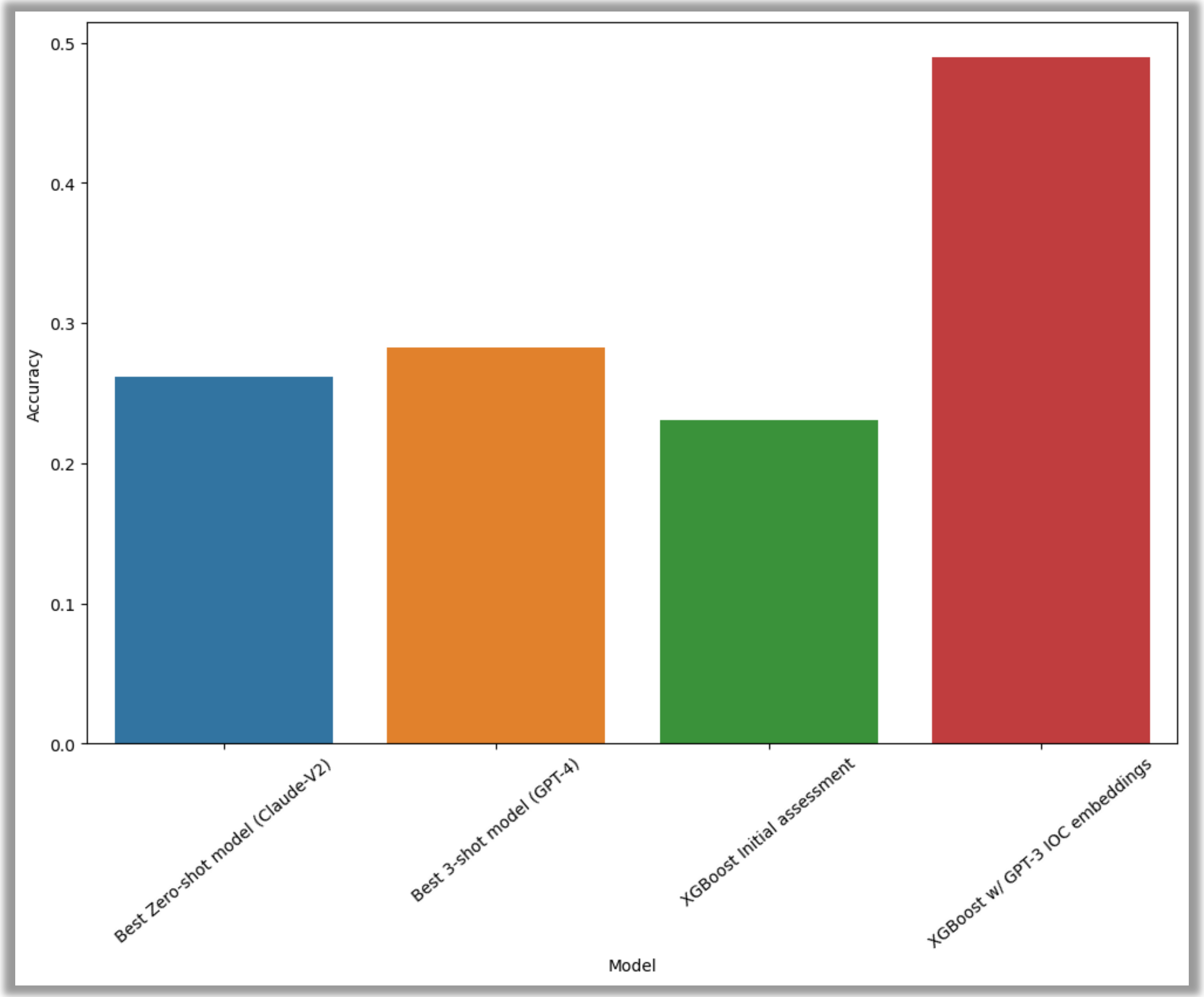
# Artifact/Incident Evaluation



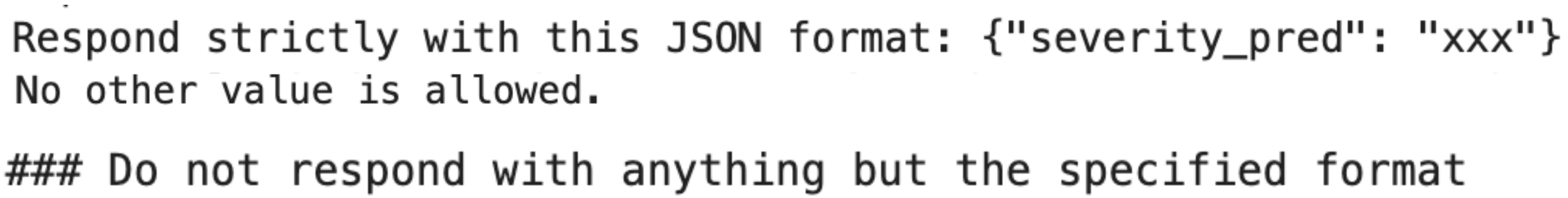
```
You are a helpful cybersecurity incident investigation expert that classifies incidents according to their severity level
given a set of detections per incident.
         Respond strictly with this JSON format: {"severity_pred": "xxx"} where xxx should only be either:
        - Critical,
            <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:
```

SOPHOS

# Artifact/Incident Evaluation Results

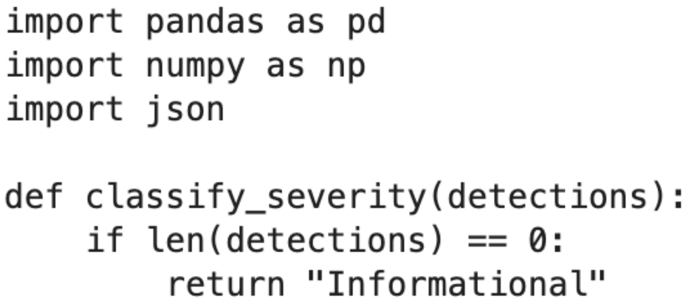


# What we said

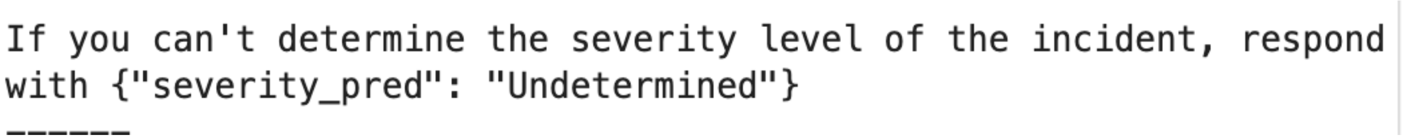


```
Respond strictly with this JSON format: {"severity_pred": "xxx"}
No other value is allowed.

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

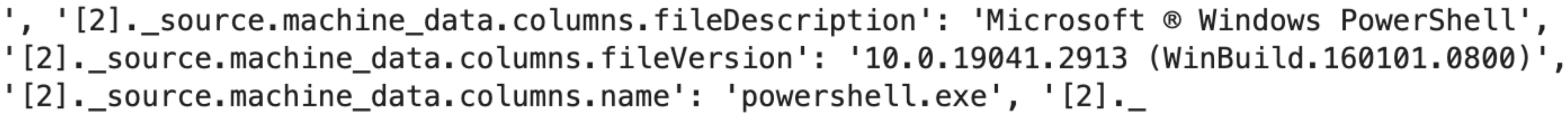
# What the models did



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

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



```
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

Want a Threat Hunting assistant? You've got one right now!

A little bit of prompt engineering will probably get you there

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?

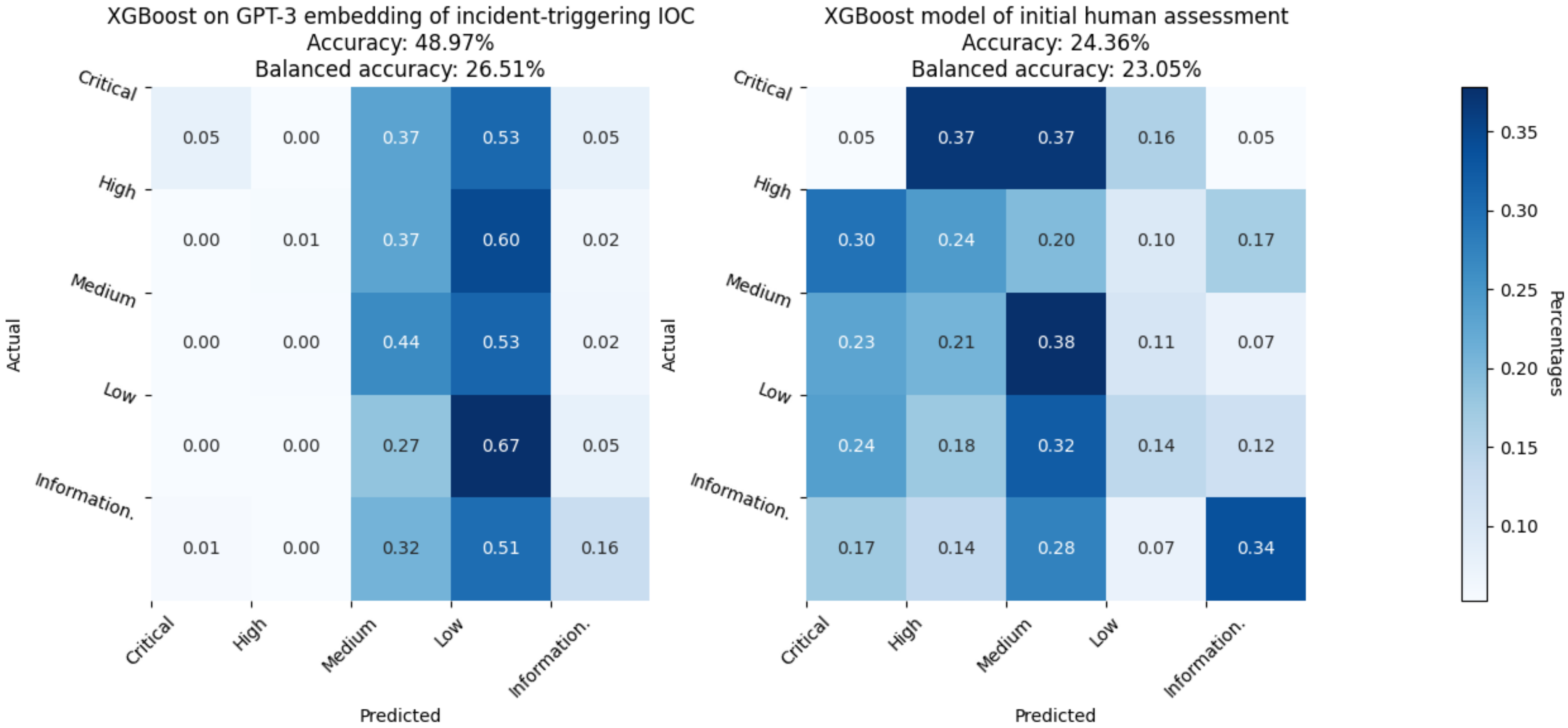
There is no good model for this right now. A good intermediate solution is to use embeddings learned by the models.

SOPHOS

Cybersecurity as a Service

# Appendix

# Artifact/Incident evaluation – confusion matrices



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

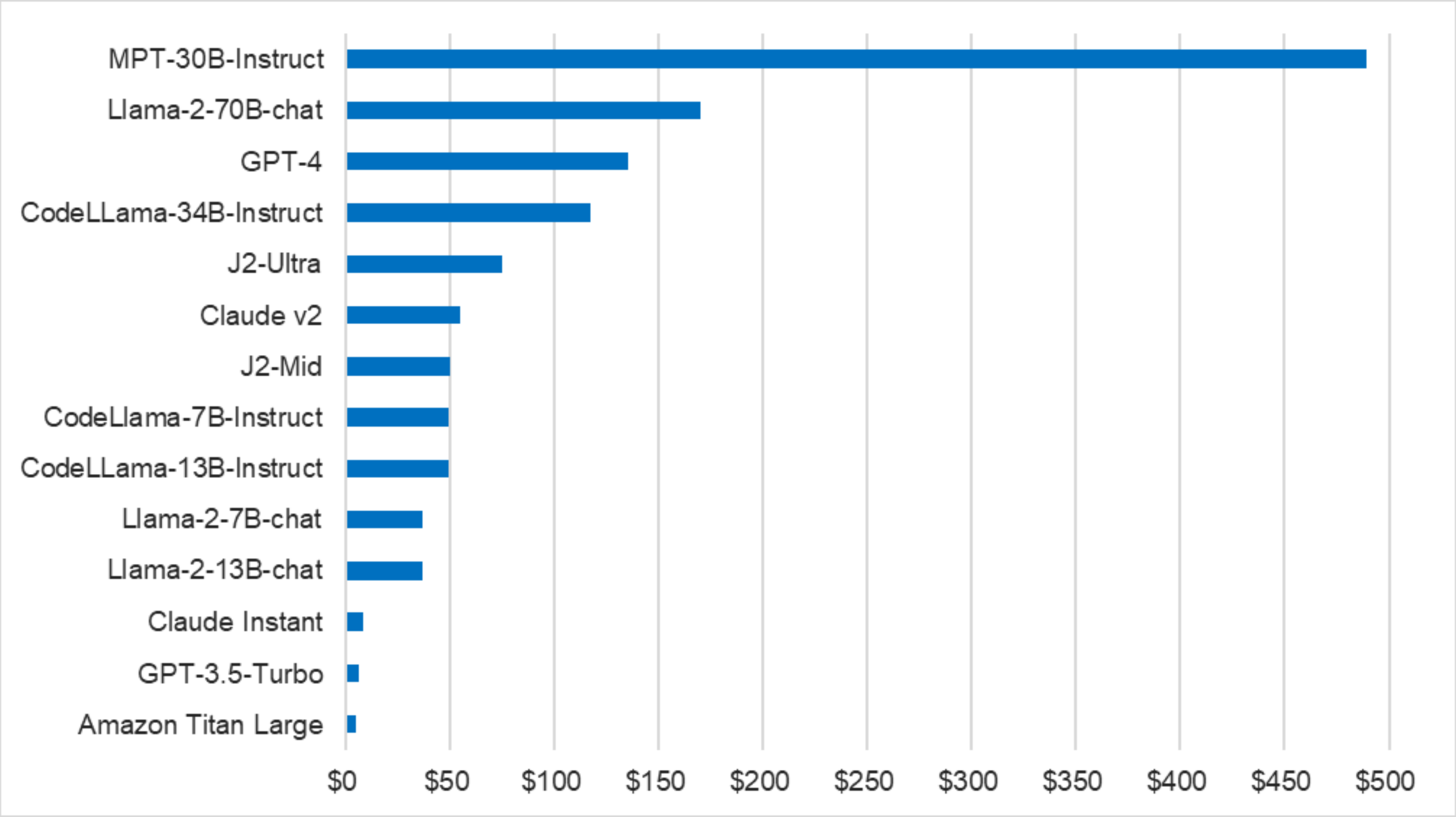
| Actual \ Predicted | Critical | High | Medium | Low | Information. |
|---|---|---|---|---|---|
| Critical | 0.05 | 0.00 | 0.37 | 0.53 | 0.05 |
| High | 0.00 | 0.01 | 0.37 | 0.60 | 0.02 |
| Medium | 0.00 | 0.00 | 0.44 | 0.53 | 0.02 |
| Low | 0.00 | 0.00 | 0.27 | 0.67 | 0.05 |
| Information. | 0.01 | 0.00 | 0.32 | 0.51 | 0.16 |

XGBoost model of initial human assessment
Accuracy: 24.36%
Balanced accuracy: 23.05%

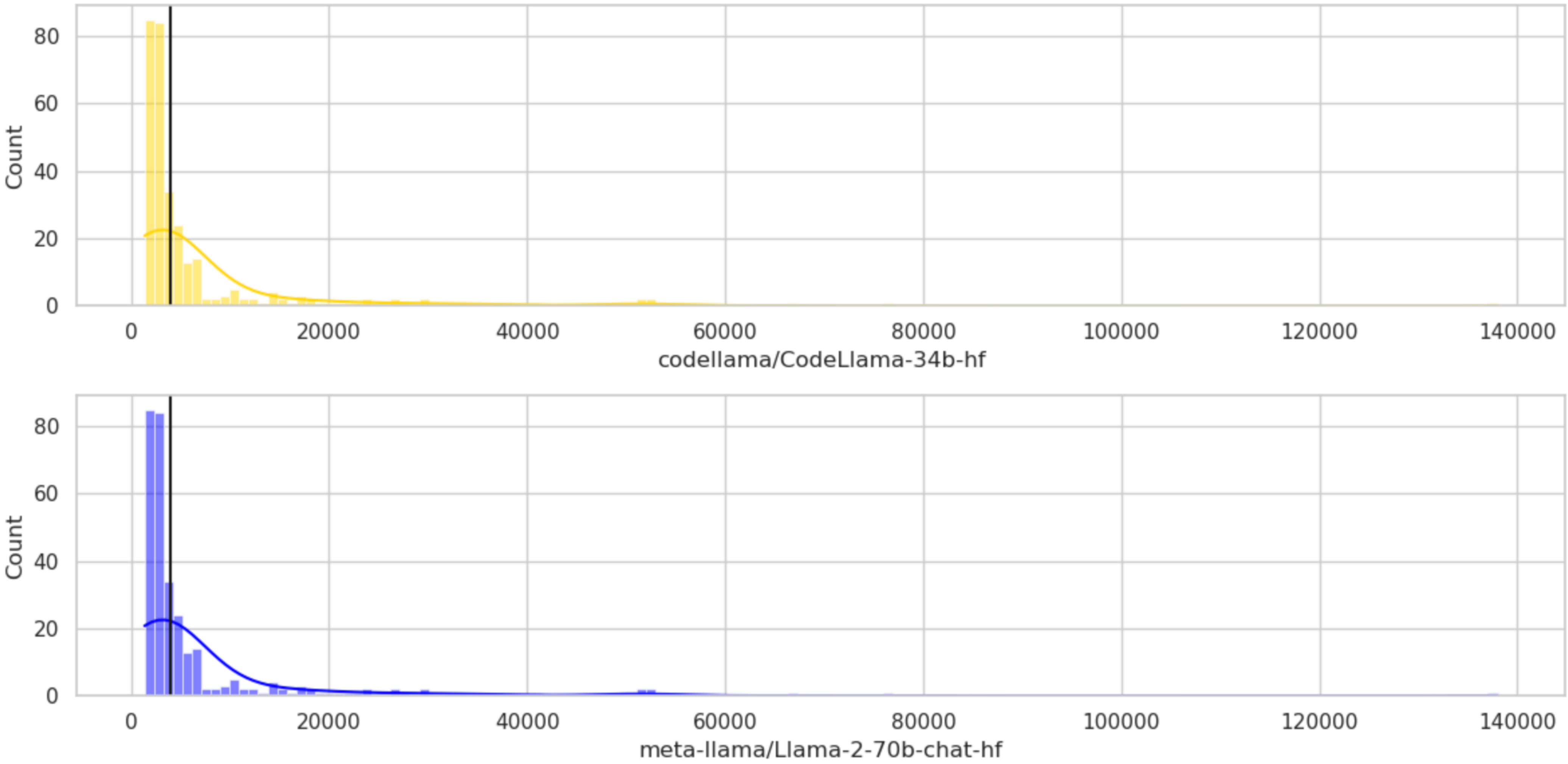
| Actual \ Predicted | Critical | High | Medium | Low | Information. |
|---|---|---|---|---|---|
| Critical | 0.05 | 0.37 | 0.37 | 0.16 | 0.05 |
| High | 0.30 | 0.24 | 0.20 | 0.10 | 0.17 |
| Medium | 0.23 | 0.21 | 0.38 | 0.11 | 0.07 |
| Low | 0.24 | 0.18 | 0.32 | 0.14 | 0.12 |
| Information. | 0.17 | 0.14 | 0.28 | 0.07 | 0.34 |

# Other important considerations

# Inference Costs



# (Useful) Context size



# Natural Language Interface – Detailed results

| Model | Rank | Accuracy |
|---|---|---|
| **GPT-4** | 1 | 0.88 |
| **Claude-Instant** | 2 | 0.85 |
| **Claude-v2** | 2 | 0.85 |
| **CodeLlama-34B-Instruct** | 2 | 0.85 |
| **GPT-3.5-Turbo** | 5 | 0.84 |
| **Llama-2-70B-Chat** | 6 | 0.83 |
| **MPT-30B-Instruct** | 7 | 0.77 |
| **CodeLlama-7B-Instruct** | 8 | 0.75 |
| **CodeLlama-13B-Instruct** | 8 | 0.75 |
| **Jurassic-2-Ultra** | 10 | 0.72 |
| **Jurassic-2-Mid** | 11 | 0.57 |
| **Llama-2-13B-Chat** | 12 | 0.53 |
| **Amazon-Titan-Large** | 13 | 0.52 |
| **Llama-2-7B-Chat** | 14 | 0.39 |

# Summarization– Detailed results

| Model | Factual Accuracy | | BERTScore (F1) | Cosine Similarity | METEOR score |
|---|---|---|---|---|---|
| | LCS-based | Levenshtein-based | | | |
| **GPT-4** | 0.7646 | 0.8162 | 0.932 | 0.9696 | 0.6955 |
| **Claude v2** | 0.6336 | 0.7235 | 0.8514 | 0.9238 | 0.3222 |
| **Claude Instant v1** | 0.5503 | 0.6488 | 0.8404 | 0.9126 | 0.282 |
| **J2-Ultra** | 0.527 | 0.6309 | 0.8359 | 0.9133 | 0.2765 |
| **J2-Mid** | 0.2862 | 0.3707 | 0.8088 | 0.8774 | 0.1934 |
| **LLaMA2 7B Chat HF** | 0.2229 | 0.3974 | 0.7782 | 0.8485 | 0.1749 |
| **LLaMA2 13B Chat HF** | 0.3786 | 0.5602 | 0.8214 | 0.8705 | 0.2363 |
| **LLaMA2 70B Chat HF** | 0.4739 | 0.6547 | 0.826 | 0.8847 | 0.2434 |
| **CodeLLaMA 7B Instruct** | 0.1189 | 0.2194 | 0.7652 | 0.8083 | 0.1039 |
| **CodeLLaMA 13B Instruct** | 0.3128 | 0.4962 | 0.7988 | 0.8488 | 0.1952 |
| **CodeLLaMA 34B Instruct** | 0.252 | 0.3746 | 0.7917 | 0.8245 | 0.164 |
| **MPT30B Instruct** | 0.2594 | 0.3783 | 0.8178 | 0.8702 | 0.1315 |
| **Amazon Titan Large** | 0.4642 | 0.5904 | 0.8418 | 0.9045 | 0.2181 |
| **GPT 3.5 Turbo** | 0.5586 | 0.6808 | 0.8624 | 0.9323 | 0.3376 |

# Artifact Evaluation – Detailed results (Zero shot)

| Model | Accuracy | Balanced Accuracy | Per-class F1 score (Critical to Informational) |
|---|---:|---:|---:|
| **GPT-4** | 0.04 | 0.2537 | [0.0245, 0.0478, 0.108, 0, 0] |
| **GPT-3.5-Turbo** | 0.26 | 0.2336 | [0.051, 0.873, 0.373, 0.270, 0.0076] |
| **Claude-v2** | 0.2 | 0.2614 | [0.035, 0.104, 0.33, 0.202, 0] |
| **Claude-Instant-v1** | 0.07 | 0.2094 | [0.013, 0.082, 0.218, 0.004, 0] |
| **J2-Ultra** | 0.49 | 0.2083 | [0.043, 0, 0, **0.66**, 0.038] |
| **J2-Mid** | 0.19 | 0.1786 | [0, 0, 0.307, 0, 0.221] |
| **Amazon-Titan-Large** | 0.007 | 0.2001 | [0.012, 0, 0, 0.002, 0] |
| **Llama-2-7B-Chat** | 0.001 | 0.063 | [0, 0, 0.267, 0.118, 0] |
| **Llama-2-13B-Chat** | 0.001 | 0.13 | [0, 0, 0, 0.67, 0] |
| **Llama-2-70B-Chat** | 0.008 | 0.05 | [0, 0.048, 0.112, 0.116, 0.283] |
| **CodeLlama-7B-Chat** | 0.02 | 0.129 | [0, 0.078, 0.118, 0.321, 0.057] |
| **CodeLlama-13B-Chat** | 0.003 | 0.128 | [0, 0.2, 0.23, 0.098, 0.077] |
| **CodeLlama-34B-Chat** | 0.016 | 0.046 | [0, 0, 0.075, 0.28, 0.28] |

# Artifact Evaluation – Detailed Results

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