Presented by Flurin Laim and Joel Müller

08/11/2025

KQLBench

Measuring LLM performance on defensive cyber security tasks

KQLBench?

- Our Idea
- How we built it
- Promising Results and Issues
- How the Future can look like

About us

Flurin Laim

Student @ HSLU

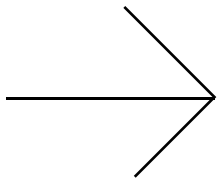
Cyber Defense Analyst @ Swisscom

Joel Müller

Student @ HSLU

Cyber Defense Engineer @ Migros

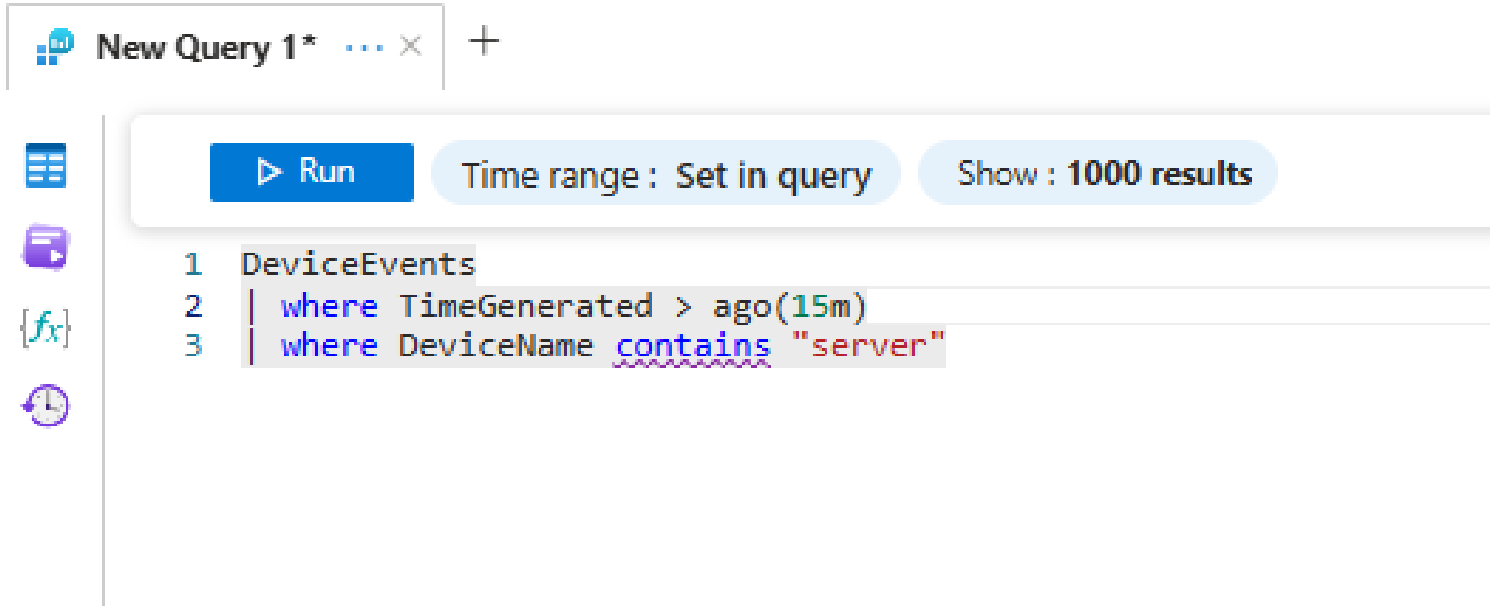
How well can LLMs detect malicious activity?



What is KQL?

(Kusto Query Language)

- Developed by Microsoft
- Mostly used for timebased Databases



Why?

- LLMs are non-determinstic
- Attackers have started using AI
- Curiosity

Current Research

Cornell University

arXiv > cs > arXiv:2404.02933

Search...

Help

Computer Science > Databases

[Submitted on 3 Apr 2024 (v1), last revised 17 Jan 2025 (this version, v4)]

NL2KQL: From Natural Language to Kusto Query

Xinye Tang, Amir H. Abdi, Jeremias Eichelbaum, Mahan Das, Alex Klein, Nihal Irmak Pakis, William Blum, Daniel L Mace, Tanvi Raja, Namrata Padmanabhan, Ye Xing

Data is growing rapidly in volume and complexity. Proficiency in database query languages is pivotal for crafting effective queries. As coding assistants become more prevalent, there is significant opportunity to enhance database query languages. The Kusto Query Language (KQL) is a widely used query language for large semi-structured data such as logs, telemetries, and time-series for big data analytics platforms. This paper introduces NL2KQL an innovative framework that uses large language models (LLMs) to

Satya Nadella • 3.+

Chairman and CEO at Microsoft

2 Wochen • Bearbeitet •

+ Folgen ···

Introducing a new open source benchmarking tool to measure AI for cybersecurity, grounded in real world scenarios. This is important work as we evaluate how well AI systems can reason to protect against cyberattacks.

Read the blog: https://lnkd.in/g4cTWH-m

And here's the GitHub repo: https://lnkd.in/g5vtSc5J

Übersetzung anzeigen

Microsoft raises the bar: A smarter way to measure AI for cybersecurity | Microsoft Security Blog

microsoft.com

2.431 114 Kommentare · 302 Reposts

Phi-4-mini

Llama 4 Scout

GLM-4.6

o3

Grok 3 mini

Sonnet 4.5

o1

GPT-5-nano

o3-mini

GPT-5-high

GPT-5-mini

DeepSeek V3.1

GPT-4.1-mini

GLM-4.6

GPT-5

DeepSeek V3

GLM-4.5

GPT-4.5

GPT-4.1-nano

Sonnet 4

MiniMax-M2

Qwen3

o3-high

GPT-4o

Opus 4.1

Sonnet 3.5

Kimi K2

Gemini 2.5 Pro

Grok 4

Llama 4 Maverick

GPT-OSS

Gemma 3

Haiku 3.5

Phi-4

GPT-5 Codex

Humanity's Last Exam

GPQA Diamond

MMLU-Pro

SciCode

SWE Bench

MMMU Pro

LiveCode

Terminal Bench

AIME 2025

AgentBench

ARC-AGI-2

JailbreakBench

HumanEval

Aider Polyglot

SafetyBench

GSM8K

BFCL

ARG-AGI-1

GRIND

DS-1000

PopQA

DROP

LongBench

MATH

MT-Bench

SWE-Bench Pro

GAIA

BEIR

ARC-AGI-3

FEVER

KQLBench

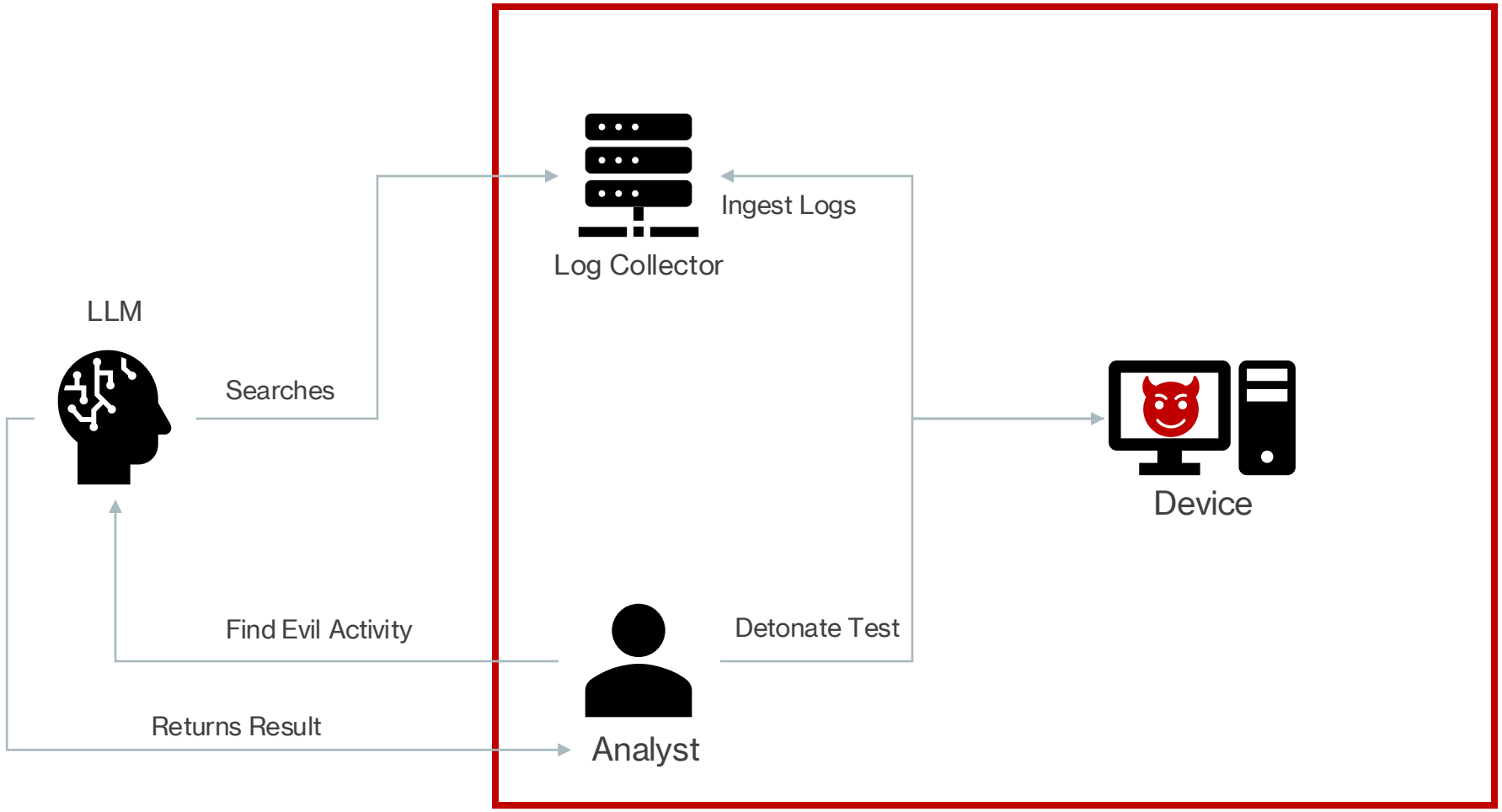
MultiRC

APPS

MBPP

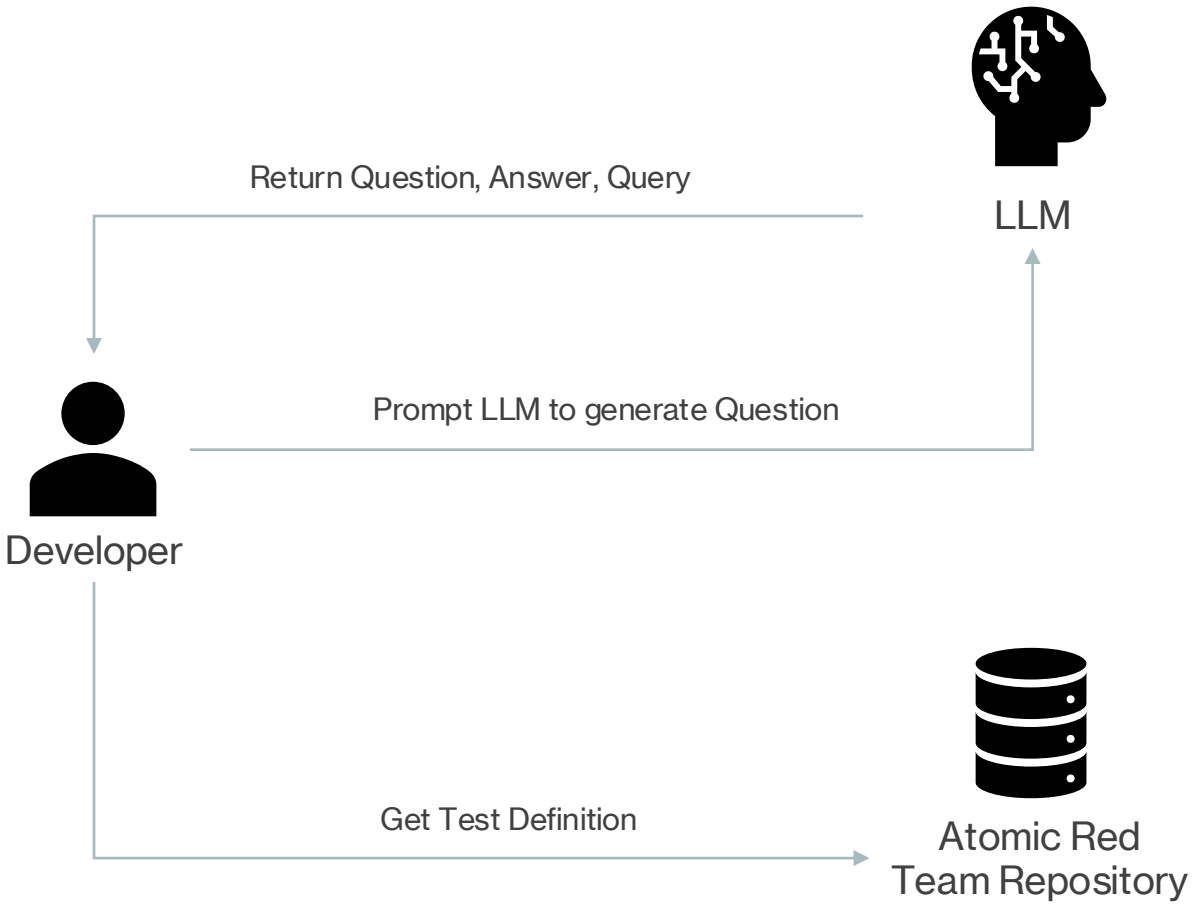
PIQA

Our Idea

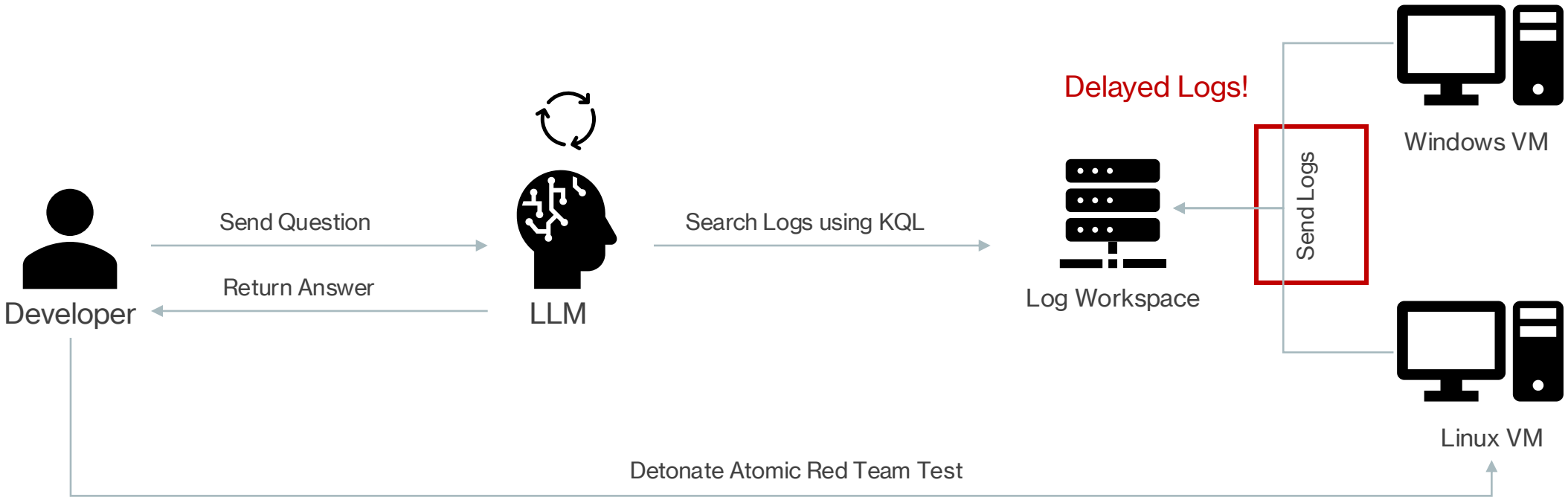


Explain Idea

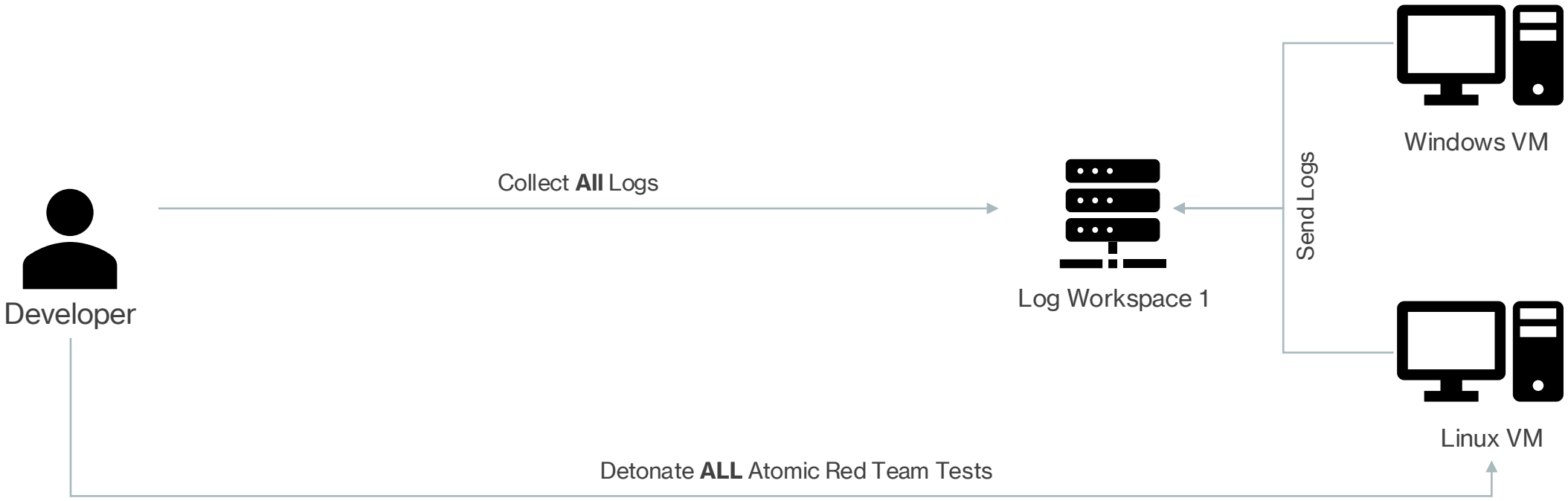
Our Idea


LLM
Return Question, Answer, Query
Prompt LLM to generate Question
Developer
Get Test Definition
Atomic Red
Team Repository

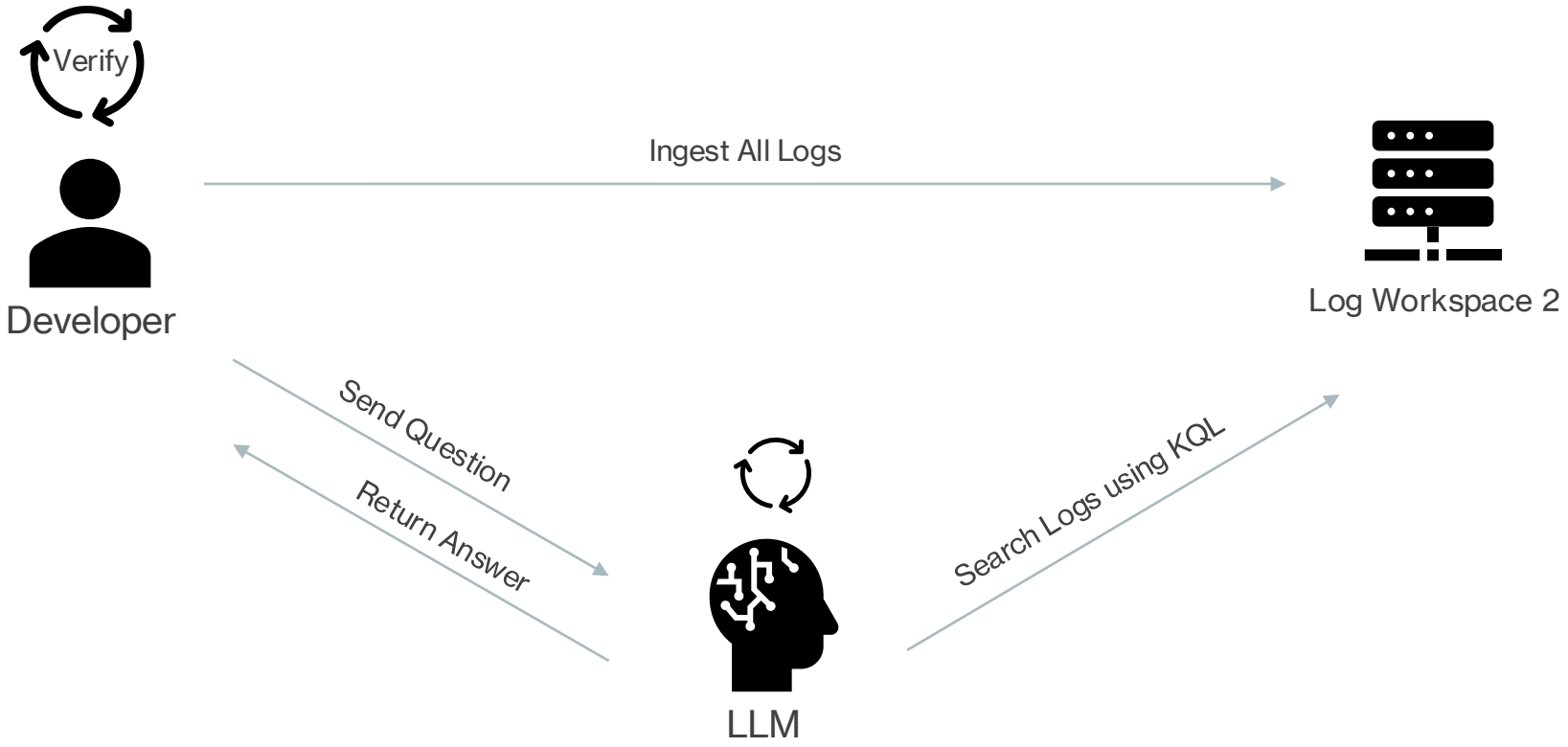
Setup


Delayed Logs!
Windows VM
Send Logs
Send Question
Search Logs using KQL
Return Answer
Developer
LLM
Log Workspace
Linux VM
Detonate Atomic Red Team Test

Setup


Developer
Collect **All** Logs
Log Workspace 1
Send Logs
Windows VM
Linux VM
Detonate **ALL** Atomic Red Team Tests

Setup


Verify
Developer
Ingest All Logs
Log Workspace 2
Send Question
Return Answer
LLM
Search Logs using KQL

Setup

Context: Brief scenario explaining what happened on the host and why it matters

Objective: What the LLM should detect or prove from the data

Technical details:

- command_analysis: How the attacker executed the action at the command level
- detection_approach: How to find this activity in logs or telemetry

Prompt: The exact question text presented to the LLM

Answer: The expected short answer that identifies the key utility or artefact

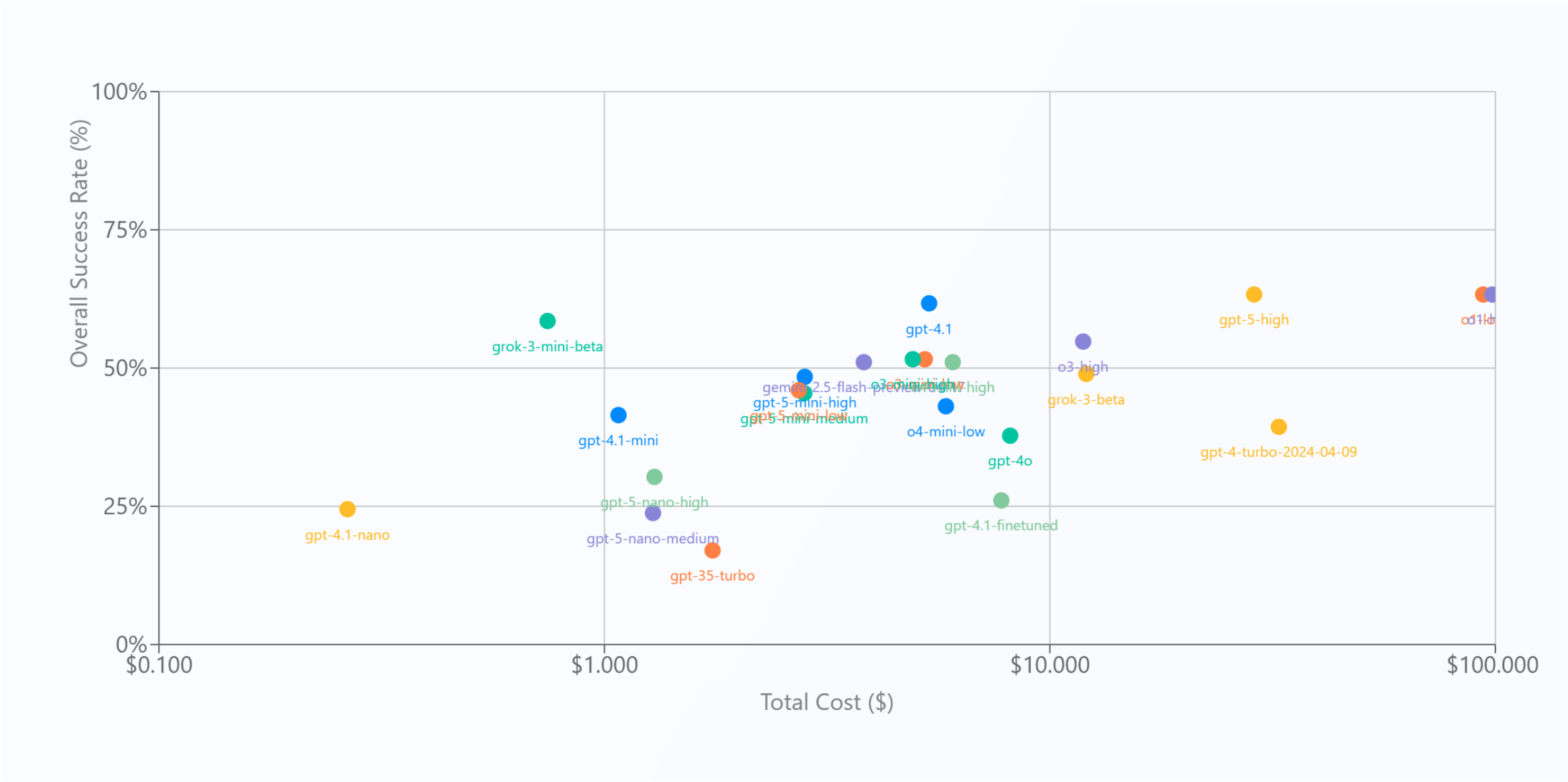
Validation query: The query used to retrieve the events for validation

Validaton Results: Output of the validation query

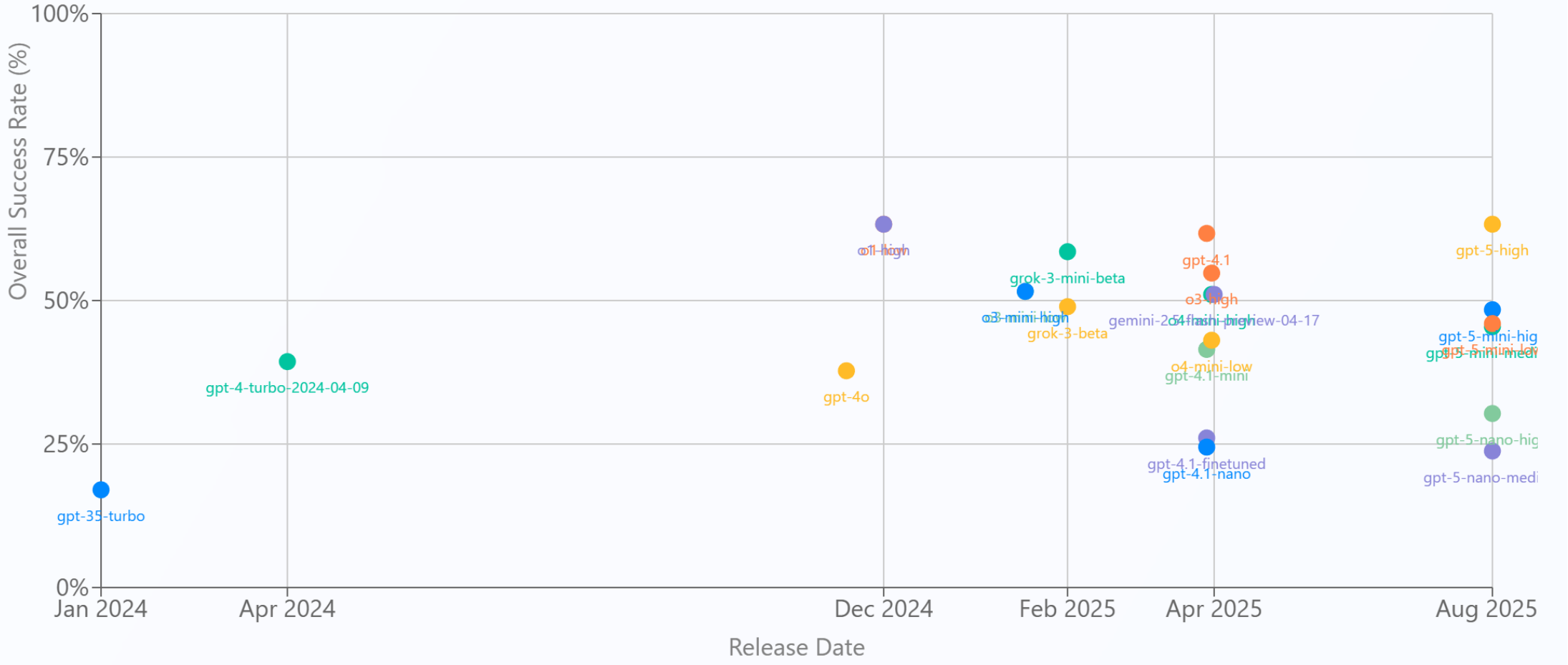
Three-Step Human Validation

236 Questions → 188 Questions

Results


100%
75%
50%
25%
0%
Overall Success Rate (%)
$0.100
$1.000
$10.000
$100.000
Total Cost ($)
grok-3-mini-beta
gpt-4.1
gpt-5-high
o3-high
grok-3-beta
gpt-5-mini-high
gpt-4.1-mini
o4-mini-low
gpt-4o
gpt-4-turbo-2024-04-09
gpt-5-nano-high
gpt-4.1-finetuned
gpt-4.1-nano
gpt-5-nano-medium
gpt-35-turbo

Results


100%
75%
50%
25%
0%
Overall Success Rate (%)
Jan 2024
Apr 2024
Dec 2024
Feb 2025
Apr 2025
Aug 2025
Release Date
gpt-35-turbo
gpt-4-turbo-2024-04-09
gpt-4o
o1-high
grok-3-mini-beta
o3-mini-high
grok-3-beta
gpt-4.1
o3-high
gemini-2.5-flash-preview-04-17
o4-mini-low
gpt-4.1-mini
gpt-4.1-finetuned
gpt-4.1-nano
gpt-5-high
gpt-5-mini-hig
gpt-5-nano-hig
gpt-5-nano-medi

Issues

T1070.003 - Indicator Removal on Host: Clear Command History

- Atomic Test #1 - Clear Bash history (rm)
- Atomic Test #2 - Clear Bash history (echo)
- Atomic Test #3 - Clear Bash history (cat dev/null)
- ...

Issues

T1070.003 - Indicator Removal on Host: Clear Command History

- Atomic Test #1 - Clear Bash history (rm)
- Atomic Test #2 - Clear Bash history (echo)
- Atomic Test #3 - Clear Bash history (cat dev/null)
- ...

Issues

Atomic Test #3 - Clear Bash history (cat dev/null)

Attack Commands:
cat /dev/null > #{history_path}

Issues

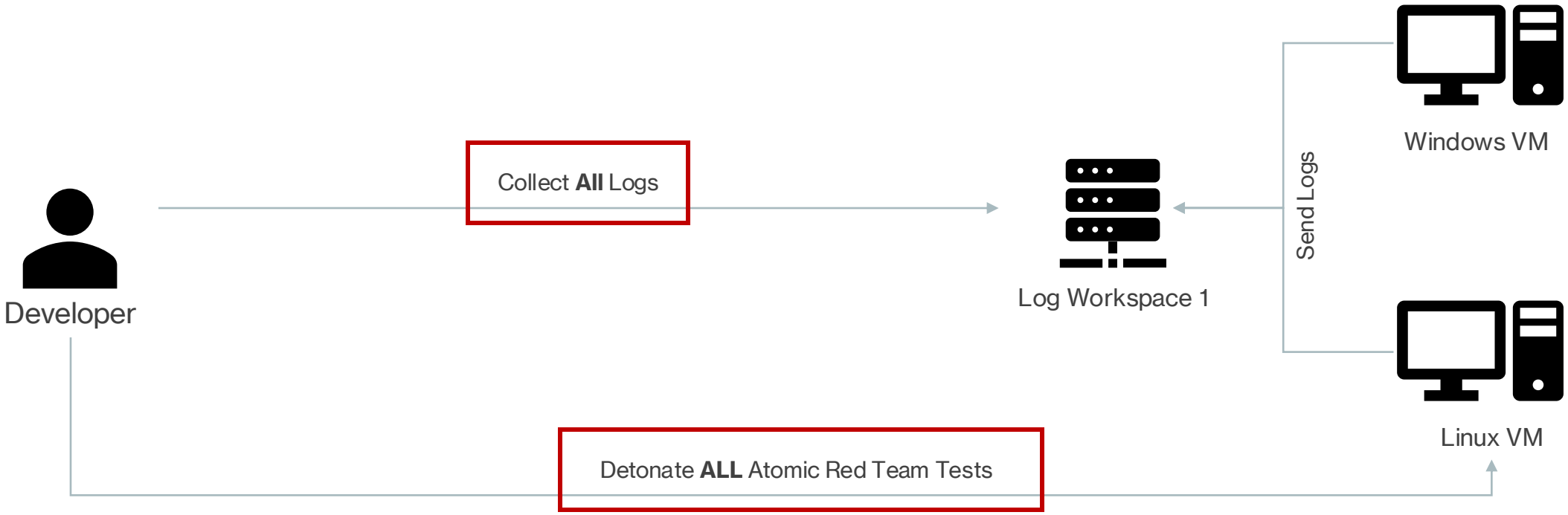
Atomic Test #2 - Clear Bash history (echo)

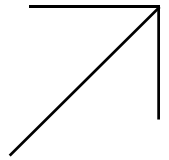
Attack Commands:
echo "" > #{history_path}

"On a Linux system, you suspect someone erased their command history. Investigate process events and determine which utility was executed to achieve this."

Answer: ln -sf /dev/null #{history_path}

Issues


Collect **All** Logs
Developer
Log Workspace 1
Send Logs
Windows VM
Linux VM
Detonate **ALL** Atomic Red Team Tests



"On a Linux system, you suspect someone erased their command history by linking the history file to /dev/null. Investigate process events and determine which utility was executed to achieve this."

Answer: ln -sf /dev/null #{history_path}

Future

- Agent Framework
- Internal adoption
- Better and cheaper Models
- More Research

Questions

kqlbench.com