

Kimo Bumanglag Threat Intel

Joseph Millman Detection and Response

A Little Less Malware, A Little More Context

Using AI to detect malicious activity



01 Introduction 02 Case studies 03 Adding context 04 Prompt engineering 05 Conclusion

We use our AI models for multiple security use case ...to reduce human burden

...to increase velocity of work

MODEL CITIZENS: OPTIMIZING THREAT ANALYSIS WITH AI-POWERED TEAMS

In this session, we'll show how we're using an agentic approach that leverages specialized assistants—including a junior analyst, a librarian, and a note-taker—to streamline investigative workflows and ensure comprehensive and structured note taking without critical information being missed. We will also share findings from our recent investigations, demonstrating how this technology empowers defenders, enabling them to detect and respond to threats more effectively than ever before.

CYBERWARCON

Intel Assistant

Plans, executes, and documents threat intelligence investigations.



LLMs at the Core: From Attention to Action in Scaling Security Teams Fotis Chantzis & Paul McMillan @PaulM (witter/infosec.exchange) @ithligore (witter)



Triage, route, and respond to common security questions.

DARPA Announces an Al Cyber Initiative, Live at DC 32 and DC 33 Piday at 14:30 in Track 2

20 minutes Moderator: Perri Adams DARPA Ak/CC Program Manager Michael Sellitto Head of Beopolitics and Security Policy, Anthropic Heather Adkins Vice President of Security Engineering, Boogle Vijay Bolina Chief Informa Security Officer & Head of Cybersecurity Research. Booale DeeMind

Dave Weston Vice President of Enterprise and OS Security, Microsoft

Matthew Knight Head of Security, OpenAl

Omkhar Arasaratnam General Manager, Open Source Security Foundation (OpenSSF)

DARFA's Al Cyber Challenge program manager, Ferri Adams, is joined by collaborators from Anthropic, Google, Google, DeepMind, OpenAl and the Open Source Security Foundation to share insights about the upcoming competition and discuss the software security challenges facing the commercial sector and open-source community.

Vulnerability management

Identify and patch vulnerabilities in software.

Can we use an AI model to augment our detection strategy?



Hypothesis: LLMs will be able to assess command line activity and surface malicious actions.





LLMs can detect malicious activity on Macs

LLMs perform better with more context

LLMs perform better with explicit instructions



2 macOS VMs

3 users (1 technical, 1 non-technical manager, 1 non-technical IC)

Benign activity (2472 total)

User simulation (Chrome, Mail, etc)

Background macOS activity (mediaanalysisd, mdworker, etc) 3 infostealers

1 loader

4 RATs

1 insider

1 ransomware

1 backdoor



 basename /Applications/Safari.app
 dscl /Local/Default -authonly michael.scott
 sh -c osascript -e 'display dialog "To launch the application, you need to update the system settings Please enter your password."
 sh -c sw_vers "group_leader.pid": 3, "hostname": "scr-office-imac", "suspcious_score": 10, "verdict": "suspicious", "mitre_tag": "credential_access", "analysis": "This sequence of commands..."

01 Introduction OP Case studies 03 Adding context 04 Prompt engineering 05 Conclusion

Stealers

Verdict	Suspicious
Score	9
Tags	["defense_evasion", "execution", "persistence"]
Analysis	The sequence of commands executed on the host 'scr-office-imac.local' by the user 'michael.scott', who is a Regional Manager, is highly suspicious. The process chain begins with an installer being executed from a mounted volume, which is unusual for a management role. Following this, there are multiple instances of base64 encoded commands being decoded and executed, which is a common technique used for obfuscating malicious activities . The decoded command involves downloading a file from an external URL 'https://megantic.online/te/54325432', which is then saved to the user's Library directory. This is followed by changing the file permissions to executable, removing extended attributes , and forcefully signing the file with codesign. These actions are indicative of an attempt to execute potentially malicious software while bypassing macOS security mechanisms . The use of 'osascript' to execute shell scripts further suggests an attempt to evade detection. Given the user's role, such technical and potentially harmful activities are highly anomalous and warrant immediate investigation.



Moonlock Lab @moonlock_lab

3/3: We found one more #macOS #loader on VT with the same script, but no file path errors. It also sets execute permissions using chmod and clears the com.apple.quarantine attribute with xattr -cr to prevent Gatekeeper from flagging the file as potentially harmful.

A										
1014							0			
							\odot		-	
N										
<u>۱</u>										
/										
							ABAR DOMANT	nya militari annany mpicati padasa ina badan pacaditari a <mark>ntana bab</mark> a		
							Reinstein () all			
								1 Martin (San) Panikana (Bittel na Hada (Bittel na Hada (Bittel na Hada (Bitte		
	an annual an annual an annual an an							 And A solution of the second se		
	C understand	APPLIEVS		descare			Retariation in			
	S Understand	Anabit	() en	identected						
						AND IN THE REPORT OF THE REPOR				
						Presses and service actives Presson Created				
		Circula .	0			C Aberginarie/beisteg/o				
	C. manual			on and the						
						C royackaya-cayanak Iteann-decide				
						Bell Lemmands				
						(hauditi-sheader) () sh ccarl a vicknarphe	et hetps://maganic.o	electrol/ADA6.thred is (Ubsrychet) A6.set: cr (Ubsrychet) A6.ceteup-for	or deep op - 1, beyrest M -	Liberghers
						th cesho Dispickey&colymula Deade-decode				
						ik casacija v bode O rovječkevitnovani David -bode?				

12:25 PM · Aug 27, 2024 · 2,120 Views

Loaders

Verdict	Suspicious
Score	9
Tags	["execution", "command_and_control"]
Analysis	The sequence of commands executed on the host scr-it-mac.local by the user Sadiq Khan, an IT Administrator, raises several red flags. The initial process involves executing an application from the Downloads folder, which is often a location for unverified software. The application attempts to display a dialog using AppleScript , which is a common technique for social engineering attacks to trick users into taking actions. The subsequent command involves using 'curl' to fetch a script from an external site 'support.docsend.site', which is suspicious as it could be used to download and execute malicious scripts. The use of 'osascript' to execute a shell script fetched from an external source is particularly concerning as it indicates potential execution of arbitrary code, which could lead to further compromise of the system. Given Sadiq Khan's role as an IT Administrator, while he might have legitimate reasons to run scripts, the execution of scripts from unverified external sources is highly unusual and risky. This activity aligns with the MITRE ATT&CK techniques of 'execution' and 'command_and_control', as it involves executing potentially malicious code and establishing a connection to an external server. Overall, the combination of downloading and executing scripts from an external source, especially from a non-standard domain, is highly suspicious and warrants further investination
5	warrants further investigation.

Group-IB Threat Intelligence 🤣 @GroupIB_TI

GROUP-IB THREAT

#Lazarus has attempted to evade detection on **#macOS** systems using a new technique - code smuggling using extended attributes



RATs

Verdict	Suspicious
Score	9
Tags	["command_and_control", "collection", "exfiltration", "defense_evasion"]
Analysis	The sequence of commands executed on the host 'scr-it-mac.local' by user 'sadiq.khan', an IT Administrator, raises significant concerns. The initial command './xpcproxy -server https://caldera.obts.training -v' suggests a potential connection to an external server, which could be part of a command and control setup. Following this, there are multiple instances of 'screencapture' commands, which indicate potential data collection activities. The use of 'base64decode' in conjunction with 'sh -c eval' commands is a common technique for obfuscating malicious scripts. Additionally, the command 'pbpaste' suggests clipboard data access, which could be used for credential harvesting. The sequence also includes 'curl' commands that upload data to an external server, indicating possible data exfiltration. The presence of commands like 'find' and 'grep' targeting specific file types and applications, including security software, suggests reconnaissance and defense evasion attempts. The combination of these activities, especially given the user's role, is highly suspicious and indicative of potential malicious behavior.

Operations	obts - 29 decisions	yesterday 01:22:28	Download Re	eport 📋 Dele	
obts Download	Graph SVG				
	scr-office-imac.loca	al scr-it	-mac local		F
+ Manual Command	+ Potential Link	running Operation Det	ails Y Filters	O base64	bfuscator:
Time A Status Ran Status	∧ Ability ∨ Name ∨	Tactic 🔶 Aç	gent 🗘 Host 🗘	pid 🔶 Lin	k Command
11/21/2024, 8:22:28 O suc PM EST	Screen Cess Capture	collection zp	scr- ovxyc office- imac.local	11300 V	iew Command
11/21/2024, 8:22:28 O suc PM EST	Screen Cess Capture	collection zjr	nuyo scr-it- mac.local	12956 V	iew Command

Insider threats

	Non-technical user	Technical user
Verdict	Benign	Benign
Score	3	1
Tags	["discovery"]	
Analysis	The sequence of commands executed on the host 'scr-office-imac.local' by the user 'michael.scott', who is a Regional Manager in the Management department, involves several directory service commands. These commands are typically used for querying and managing directory services on macOS, such as Active Directory. The commands executed include listing groups, users, and computers in the Active Directory, as well as showing directory service configuration. While these commands are technical and typically associated with IT or administrative roles , they are not inherently malicious. However, given that Michael Scott is a Regional Manager, it is somewhat unusual for him to be executing such commands unless he has a specific reason or technical background. The execution of a script named 'user_audit.sh' suggests a possible audit or review of user accounts, which could be legitimate if authorized. There is no indication of privilege escalation, data exfiltration, or other malicious activities in this command sequence. Therefore, while the activity is slightly anomalous for a management role, it does not appear to be malicious . It is advisable to verify the legitimacy of the 'user_audit.sh' script and ensure that Michael Scott has the appropriate permissions and reasons for conducting such audits	The sequence of commands executed on the host 'scr-it-mac.local' by the user 'sadiq.khan', who is an IT Administrator in the Corporate and Human Resources department, appears to be consistent with typical administrative tasks. The execution begins with a script 'user_audit.sh' being run with elevated privileges using 'sudo'. This is followed by several 'dscl' commands, which are used to query directory services for information about users, computers, and groups. These commands are common in environments where directory services are used to manage network resources and user accounts. The 'dsconfigad -show' command is used to display the current configuration of the Active Directory plugin, which is also a typical administrative task. Given the user's role as an IT Administrator , these commands are within the scope of their responsibilities and do not indicate malicious intent . The use of 'sudo' and directory service commands is expected for someone in this position, and there are no signs of data exfiltration, privilege escalation, or other suspicious activities. Therefore, the activity is assessed as benign.
i fa	appropriate permissions and reasons for conducting such addits.	40

Insider threats

	Non-technical, non-manager user
Verdict	Suspicious
Score	7
Tags	["discovery"]
Analysis	The sequence of commands executed on the host 'scr-office-imac.local' by the user 'kelly.kapoor', a Customer Service Representative, involves several directory service commands (dscl) and a directory service configuration command (dsconfigad). These commands are typically used for querying and managing directory services, such as Active Directory, which is not typically within the scope of a Customer Service Representative's role . The execution of a script located in a corporate share directory ('/Volumes/Corporate Share/it/scripts/user_audit.sh') suggests that this activity might be part of an IT audit or maintenance task. However, the role of the user does not align with performing such tasks, which raises suspicion . The commands executed are related to user and group enumeration, which could be associated with discovery activities in a malicious context. Given the user's role and the nature of the commands, this activity is anomalous and warrants further investigation to ensure it is authorized and legitimate.

Ransomware

Verdict	Benign
Score	2
Tags	
Analysis	The commands in question involve the execution of a binary named 'lckmac' with elevated privileges using 'sudo'. The user executing these commands is 'sadiq.khan', who holds the position of IT Administrator in the Corporate and Human Resources department. Given the user's role, it is plausible that they have legitimate reasons to execute administrative tasks that require elevated privileges. The process parent name '/bin/bash' and the use of 'sudo' suggest that the user is executing these commands manually, possibly for system maintenance or configuration purposes. The binary 'lckmac' is not a standard macOS utility, which could raise some concerns. However, without additional context or evidence of malicious behavior, such as network connections to suspicious domains or further suspicious command sequences, this activity alone does not strongly indicate malicious intent. Therefore, considering the user's role and the nature of the command, this activity is assessed as benign.



SentinelOne 🤣 @SentinelOne

We dubbed it NotLockBit for a reason: This latest macOS malware discovery appears to be an entirely different threat actor appropriating the name of a more notorious one. SentinelOne has discovered a set of related Mach-O samples in addition to the one reported on recently by other researchers.

In a new blog post by @philofishal, we provide indicators of compromise for this set of samples and discuss how they have changed across versions. Learn more: s1.ai/NotLockBit

SentinelOne

FROM THE FRONT LINES

New NotLockBit macOS Malware Variants Indicate Threat Actor Sharpening Its Tools

We have discovered a set of related Mach-O samples in addition to the one reported on previously.

By Phil Stokes

12:40 PM · Oct 27, 2024 · 1,386 Views

01 Introduction 02 Case studies 03 Adding context 04 Prompt engineering 05 Conclusion

How much does providing extra context improve GPT detection quality?

Hypothesis: We should see improved analysis when more context on the process and user is provided to GPT.

Adding context Low context approaches



Isolated Commands

Analyze each command individually - lowest context



Grouped Commands

Analyze commands grouped by PGID - medium context

Adding context Isolated Command Performance

Low context performance

Command	/usr/bin/syslog -s -k com.apple.message.domain com.apple.security.assessment.current_state com.apple.message.signature assessments disabled com.apple.message.signature2 devid enabled Message Gatekeeper state assessments disabled/devid enabled
Verdict	Suspicious
Score	5
Tags	["defense_evasion"]
Analysis	The command executed by the syslog process involves querying or modifying the state of Apple's Gatekeeper, which is a security feature designed to ensure that only trusted software runs on a Mac. The command uses the syslog utility to log or retrieve system messages related to the security assessment state of the system. This activity could be part of routine system administration or security monitoring to ensure that Gatekeeper is configured correctly. However, if this command is executed without proper authorization or in an unusual context, it could indicate an attempt to bypass security controls, which would be suspicious. If this command was executed by a known and trusted administrator or security tool, it is likely benign.



Malicious

Predicted label

Benign

\$

\$



Grouped Commands + Role Description

Analyze commands grouped by PGID enriched with who the user is, their job title, and department - high context Adding context Contextual performance

GPT benefits from context

- Improved context for GPT analysis
- Context aids in detecting
 - Renaming utilities
 - Obfuscated commands
 - Using obscure tooling
- Group by:
 - Time
 - Process group
 - $\circ \quad \ \ Session$



01 Introduction 02 Case studies 03 Adding context 04 **Prompt engineering** 05 Conclusion

Prompt engineering Fundamentals

Prompt Engineering Strategy

https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering

Write clear instructions

- Provide reference text
- Split complex tasks into simpler subtasks
- Give the model time to "think"
- Test changes systematically

INSTRUCTIONS

You are an expert security professional at a paper company. Your security knowledge and judgment are unparalleled. You analyze security data, and output analysis in JSON format. You always follow user instructions.

You will be given a list of processes and commands from a host that have been surfaced by an alert to analyze. Your job is to analyze the activity for malicious activity. You will provide a 1–10 score of the suspiciousness of the process group and tag the command with the corresponding MITRE ATT&CK step, if applicable. Prompt engineering Fundamentals

- Write clear instructions
- Provide reference text
- Split complex tasks into simpler subtasks
- Give the model time to "think"
- Test changes systematically

Command	echo 'c2ggLWMgJ2RtdW5nlC1yZiAvJCA+IC9kZXYvbnVsbC c=' base64decode bash
Verdict	Suspicious
Score	10
Tags	["execution"]
Analysis	The base64-encoded command decodes to 'sh -c 'dmung -rf / \ > /dev/null'', which suggests a delete operation of the entire filesystem. This is potentially destructive activity indicative of severe malicious intent.



- Write clear instructions
- Provide reference text
- Split complex tasks into simpler subtasks
- Give the model time to "think"
- Test changes systematically

ANALYSIS

- You are provided with activity from related processes. During your analysis take into account the sequence of activity in the commands and frame each command in the context of the previous commands. For example, a kubectl exec execution might be benign if the endpoint proceeds to just read a log file, but if the command is followed with a curl command to a suspicious server to run a script, this is more suspicious. Several commands might appear benign in isolation, but in aggregate could indicate malicious behavior. Analyze commands with this in mind.

- Only analyze commands that are presented in the array of commands passed in the input. Do not generate your own examples of commands. Do not return the example commands or variations on the example commands.

- Only mark verdict as "suspicious" if you are certain the activity is highly suspicious and merits review by an analyst. Keep false positives to a minimum. We have a small team and taking up time responding to false positive alerts will decrease our overall security by giving us less time to respond to and triage true positives.

- You will be provided with the role and department of the employee whose commands are being analyzed. Take the individuals role into account when analyzing the command chain. Where relevant, explain why a command may be fitting or anomalous for the user, given their role. Use this information to inform your scoring of the process group and your verdict. Highly anomalous commands for a given role should be scored as more suspicious. No users should be engaging in activity that involved setting up reverse shells, reading a high number of credentials, or exfiltrating large amounts of data.

How important is the level of detail provided in the prompt?

Hypothesis: We should see decreased performance when we utilize a barebones prompt.



INSTRUCTIONS

You will be given a list of processes and commands from a host that have been surfaced by an alert to analyze. Your job is to analyze the activity for malicious activity. You will provide a 1– 10 score of the suspiciousness of the command and a suspicous/binary verdict. These commands are all related and are all descendants of the same process.

OUTPUT

- Follow this output format
{"hostname": "hostname",
"group_leader.pid": "group_leader.pid",
"analysis": "<your own analysis of 400 or fewer words>",
"verdict": <benign_or_suspicious>,
"suspicious_score": <your score of 0-10>,

Example barebones prompt

Evaluation Comparison

Detailed Prompt vs. Short Prompt

- Long prompt performs slightly better (fewer false positives)
- Short prompt is equivalent in surfacing malicious behavior
- In our experience, more detail helps in an enterprise context



Abbreviated Prompt Precision: .38 Recall: .93

01 Introduction 02 Case studies 03 Adding context 04 Prompt engineering 05 Conclusion

Takeaways

LLMs can detect malicious activity on Macs

LLMs can **augment**, not replace, your detection strategy

40 provides greater depth of analysis

4o-mini is cheaper and faster

LLMs perform better with more context

Group commands to **represent what happened** in a session

Provide information on who the employee is and what their role does

Manage token input sizes

LLMs perform better with explicit instructions

Give **GPT guidance** on how to analyze security events

Version control and evals for prompt development

Structured Outputs improve compatibility into existing tooling (SIEM, SOAR)

Conclusion What's next

The Future

- Models continue to improve
 - Context window increases
 - Costs lower
- Fine-tuning
 - We used base GPT4 family models
 - Explore fine-tuning for detection-specific tasks
- Retrieval
 - Provide access to internal data to make detection more relevant to our environment
- Search
 - Update context with current web sources
- Reasoning models
 - "Think harder" on the problem and develop a response plan
- Apply to other telemetry
 - Application logs
 - Network data



The model's creative writing ability has leveled up–more natural, engaging, and tailored writing to improve relevance & readability.

It's also better at working with uploaded files, providing deeper insights & more thorough responses.



X

Thank you

Contact us

Kimo Bumanglag kimo (at) openai.com Joseph Millman joseph (at) openai.com Shout outs

Bob, Matt, Flossman, Harold, Akshay, McGrew