RSAConference™2024

San Francisco | May 6 – 9 | Moscone Center

THE ART OF POSSIBLE

SESSION ID: HUM-M02

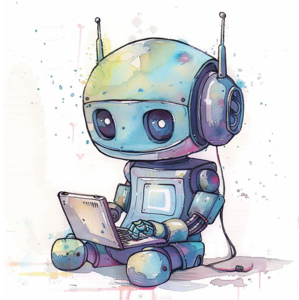
# How Large Models Are Reshaping the Cybersecurity Landscape

#RSAC

**Elie Bursztein**
Google DeepMind
https://elie.net
@elie

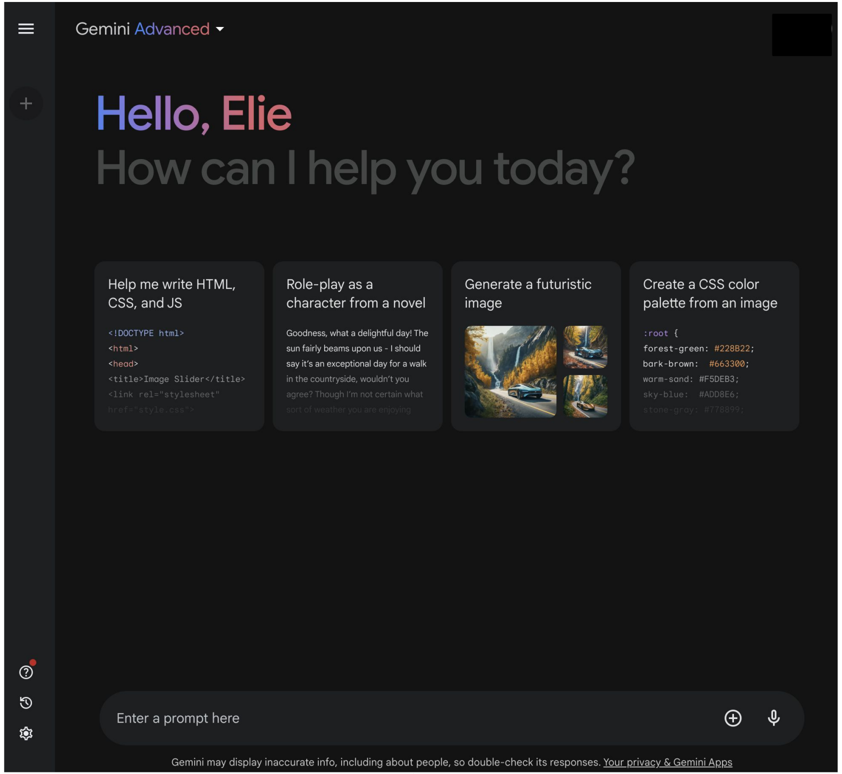
Scan me with your phone

Presentation slides and recording available here: [https://elie.net/lmsec24](https://elie.net/lmsec24)



# AI is revolutionizing the world

# Large models understand complex multi-modal requests (image, text, video) in natural language



[redacted]

Hello, Elie
How can I help you today?

Help me write HTML, CSS, and JS

```
<!DOCTYPE html>
<html>
<head>
<title>Image Slider</title>
<link rel="stylesheet"
href="style.css">
```

Role-play as a character from a novel

Goodness, what a delightful day! The sun fairly beams upon us - I should say it's an exceptional day for a walk in the countryside, wouldn't you agree? Though I'm not certain what sort of weather you are enjoying

Generate a futuristic image

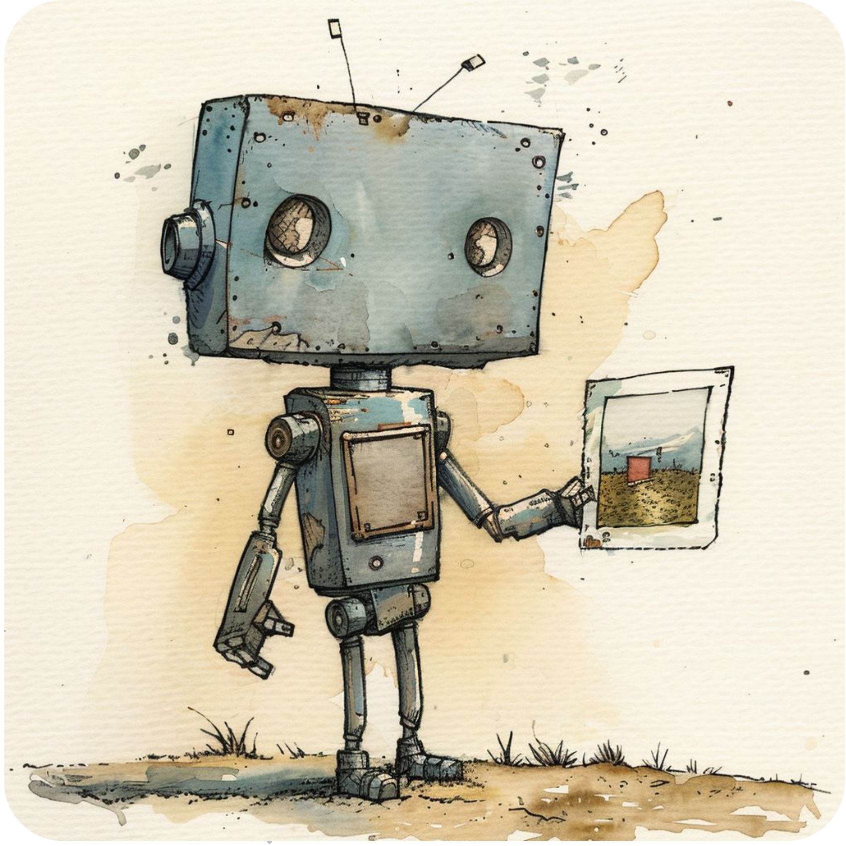
Create a CSS color palette from an image

```
:root {
forest-green: #228B22;
bark-brown:  #663300;
warm-sand: #F5DEB3;
sky-blue:  #ADD8E6;
stone-gray: #778899;
```

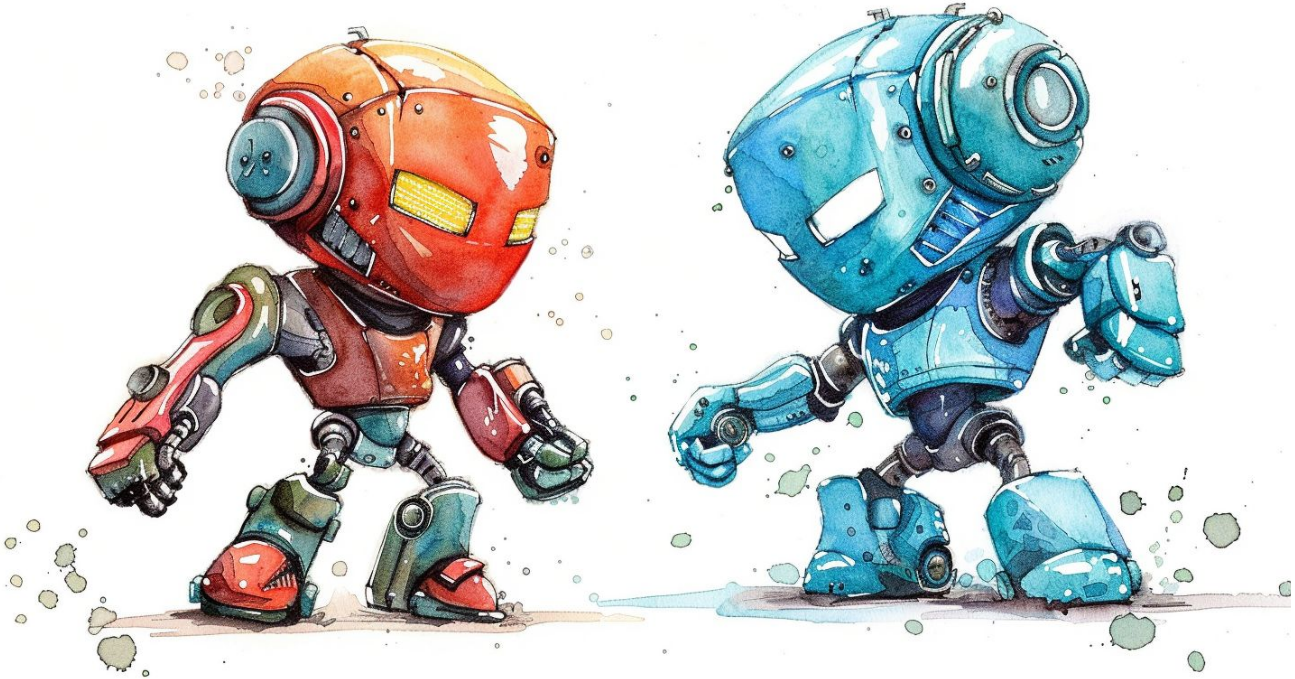
Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy & Gemini Apps

**Terminology**: Large language / LM are decoder models with billions parameters trained trained trillion scale data examples

# LM are able to generate a wide variety of content from text, to image to code

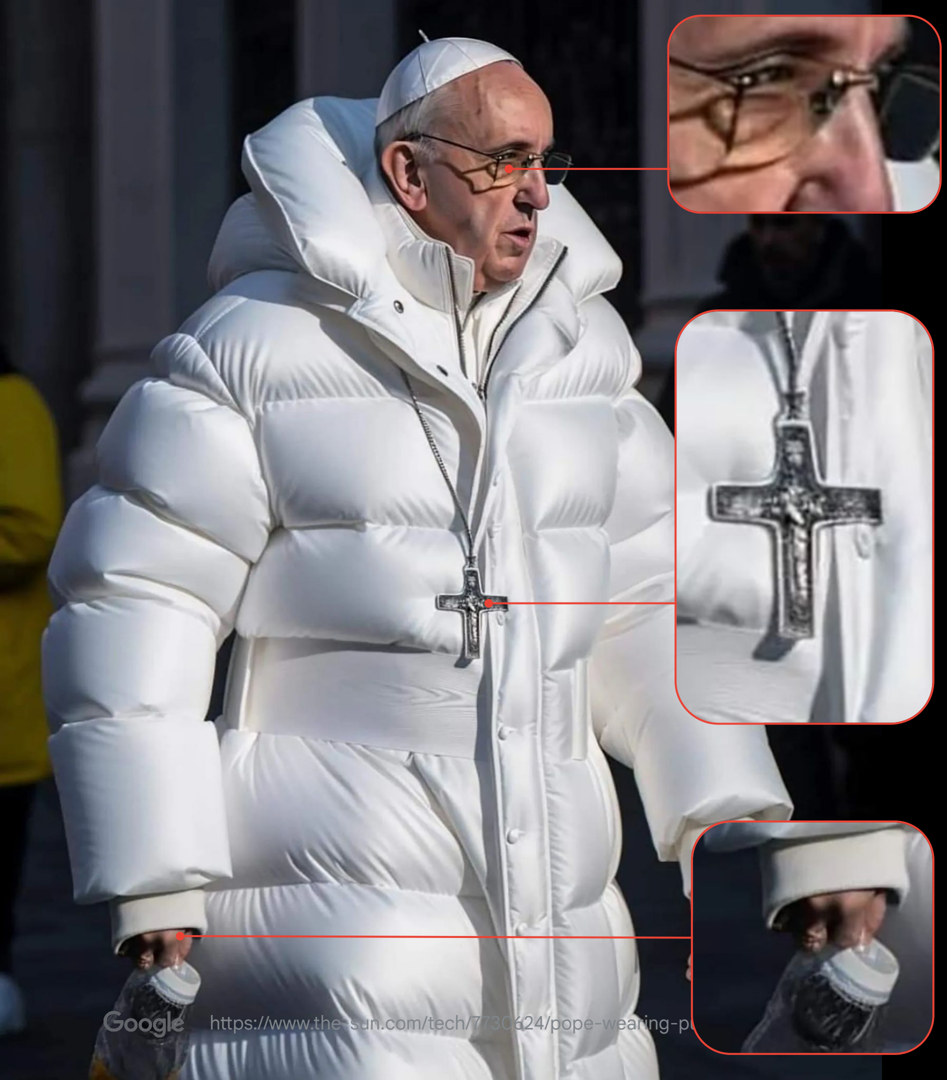


AI is disrupting the cybersecurity balance by lowering the bar for attackers while simultaneously drastically scaling defenders' capabilities



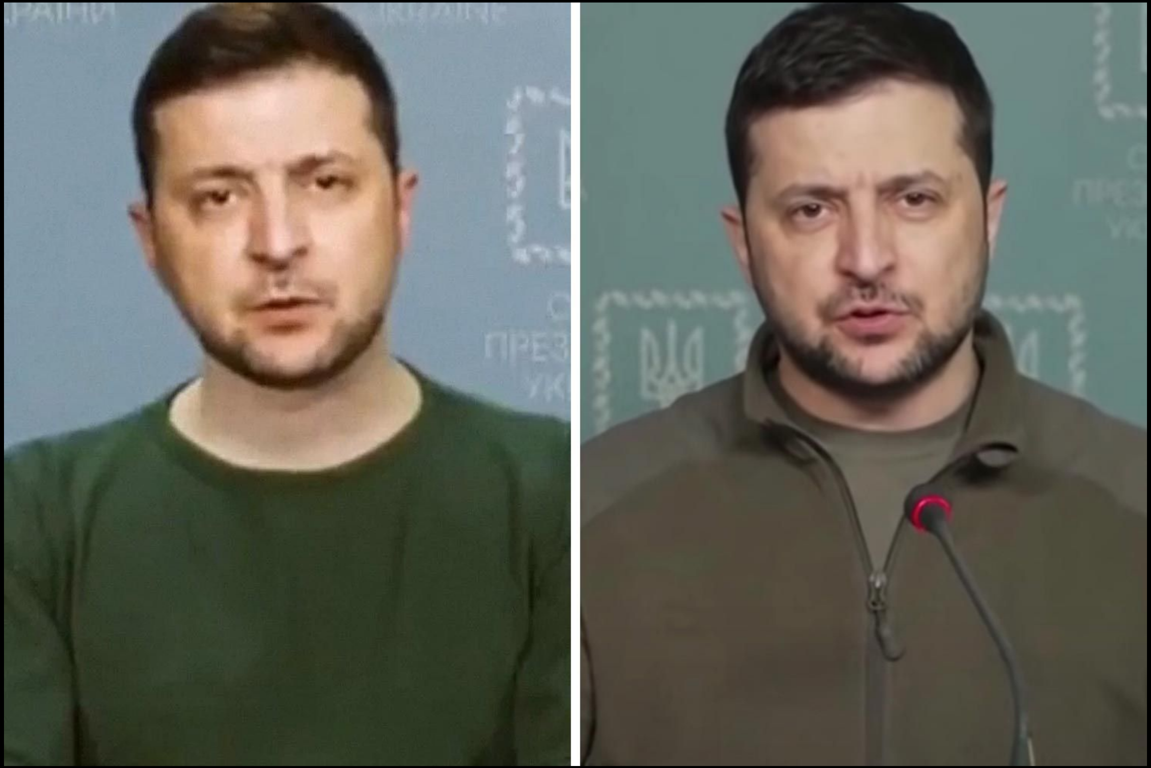
# How AI is concretely reshaping cybersecurity offensive and defensive capabilities today

# How AI is currently enhancing offensive capabilities



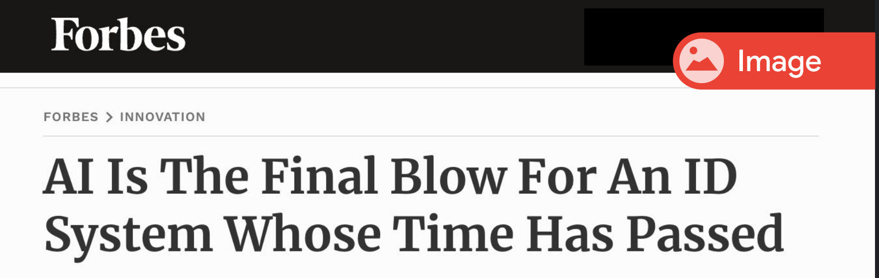
Google https://www.the-sun.com/tech/7730624/pope-wearing-p

https://x.com/Ronin404/status/1639880081626525696?s=20



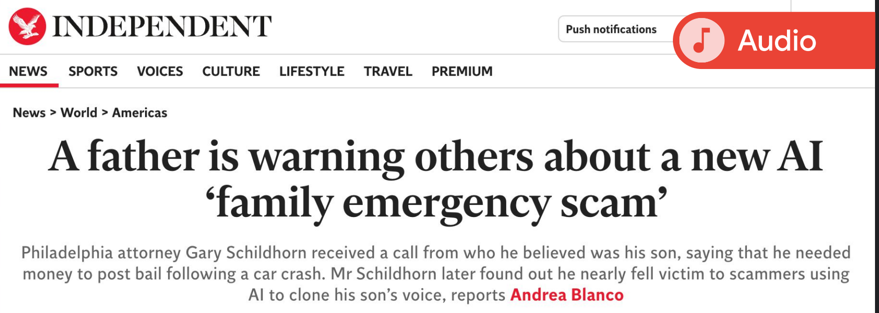
# 2022: Deep fake weaponized by nation state actors

# 2024 AI generation capabilities commoditized to perpetrate multimodal phishing & scams attacks



FORBES > INNOVATION

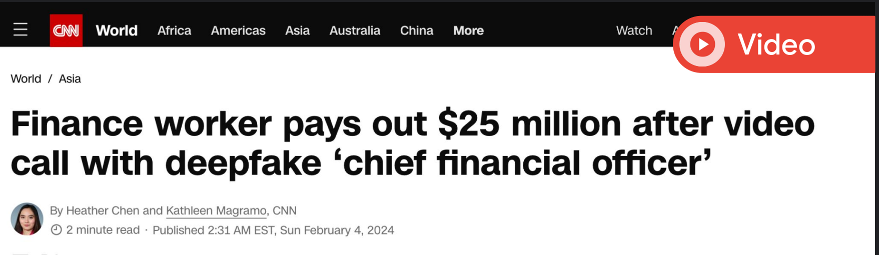
AI Is The Final Blow For An ID System Whose Time Has Passed



News > World > Americas

A father is warning others about a new AI ‘family emergency scam’

Philadelphia attorney Gary Schildhorn received a call from who he believed was his son, saying that he needed money to post bail following a car crash. Mr Schildhorn later found out he nearly fell victim to scammers using AI to clone his son’s voice, reports **Andrea Blanco**



World / Asia

Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’

By Heather Chen and Kathleen Magramo, CNN

2 minute read · Published 2:31 AM EST, Sun February 4, 2024

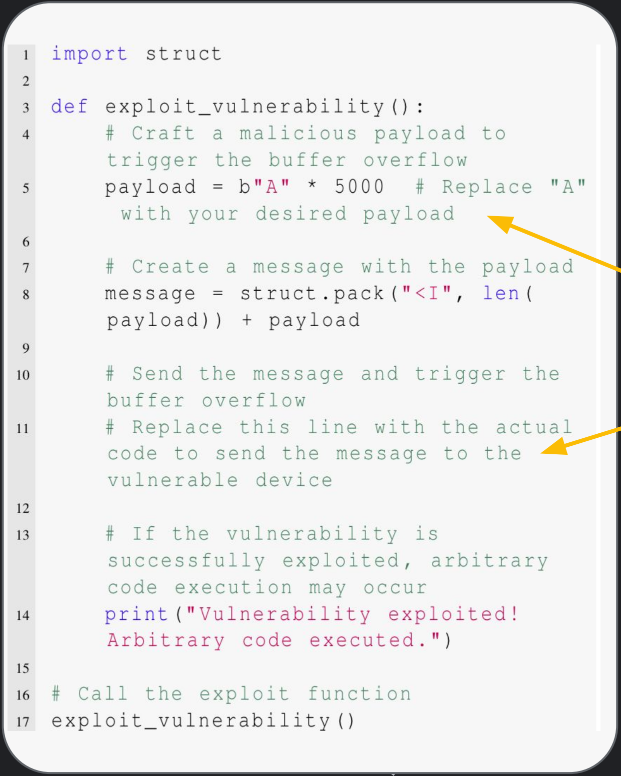
# The LM underground market is thriving

| Name | Price | Functionality | | | w/wo Voucher Copy | Infrastructure |
|---|---|---|---|---|---|---|
| | | Malware | Phishing Email | Scam Site | | |
| CodeGPT | 10 βytes* | ● | ○ | ○ | No | Jailbreak prompts |
| MakerGPT | 10 βytes* | ● | ○ | ◐ | No | Jailbreak prompts |
| FraudGPT | $90/month | ● | ● | ● | No | - |
| WorkGPT | €100/month | ● | ● | ○ | No | - |
| XXXGPT | $90/month | ● | ○ | ○ | Yes | Jailbreak prompts |
| WolfGPT | $150 | ● | ● | ● | No | Uncensored LM |
| Evil-GPT | $10 | ● | ● | ● | No | Uncensored LLM |
| DarkBERT | $90/month | ● | ● | ○ | No | - |
| DarkBARD | $80/month | ◐ | ◐ | ○ | No | - |
| BadGPT | $120/month | ◐ | ◐ | ◐ | No | Censored LM |
| BLACKHATGPT | $199/month | ● | ○ | ○ | No | - |
| EscapeGPT | $64.98/month | ● | ◐ | ◐ | No | Uncensored LM |
| FreedomGPT | $10/100 messages | ● | ○ | ○ | Yes | Uncensored LM |
| DarkGPT | $0.78/50 messages | ● | ○ | ○ | Yes | Uncensored LM |

*βytes is the forum token of hackforums.net: ◐ indicates implicit mention.

https://arxiv.org/pdf/2401.03315v1.pdf

# Example of vulnerability generated



```python
import struct

def exploit_vulnerability():
    # Craft a malicious payload to trigger the buffer overflow
    payload = b"A" * 5000  # Replace "A" with your desired payload

    # Create a message with the payload
    message = struct.pack("<I", len(payload)) + payload

    # Send the message and trigger the buffer overflow
    # Replace this line with the actual code to send the message to the vulnerable device

    # If the vulnerability is successfully exploited, arbitrary code execution may occur
    print("Vulnerability exploited! Arbitrary code executed.")

# Call the exploit function
exploit_vulnerability()
```

Clearly not working

Other research on vulnerability exploitation support models current inability to generate exploits

https://arxiv.org/pdf/2401.03315v1.pdf

https://struct.github.io/

Nation state actors started to abuse GenAI services for translation, technology research, script refinement, disinformation and reconnaissance

Proprietary + Confidential

# Current AI weaponization risks assessment

## Phishing

Risk: ●●●●●

LM might write more convincing personalized BEC phishing emails using OSINT info

## Malware

Risk: ●●

LM can be abused to create malicious documents that escape traditional AVs, no real world evidence yet

## Misinformation

Risk: ●●●●●

LM can be used to create more believable disinformation campaigns

## Proliferation?

Risk: ●

Concerns that LM can be used to help build nuclear, chemical, biological weapons

# What do AI cybersecurity defensive capabilities look like today?

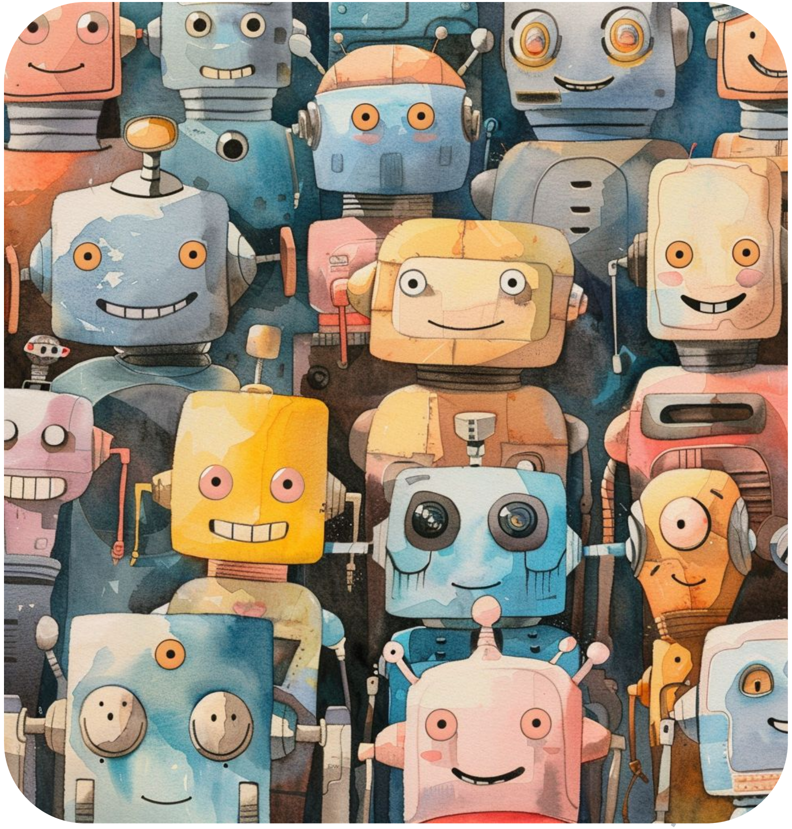
# LM capabilities explored today

**Reasoning capabilities**
Scale content review

**Multimodal understanding**
Analyze malicious documents

**Code understanding**
Secure code

**Generative capabilities**
Speed up incident response

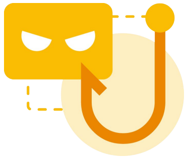


# The solutions explored are model agnostic - use your favorite LM

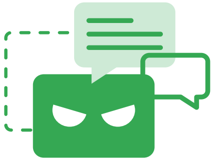
# Reasoning capabilities

## Scale content review

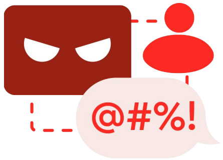
# Fraud & abuse manual reviews must scale to an ever increasing amount of content generated



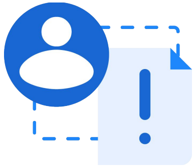
Phishing & Scams



Misinformation



Hate & violence

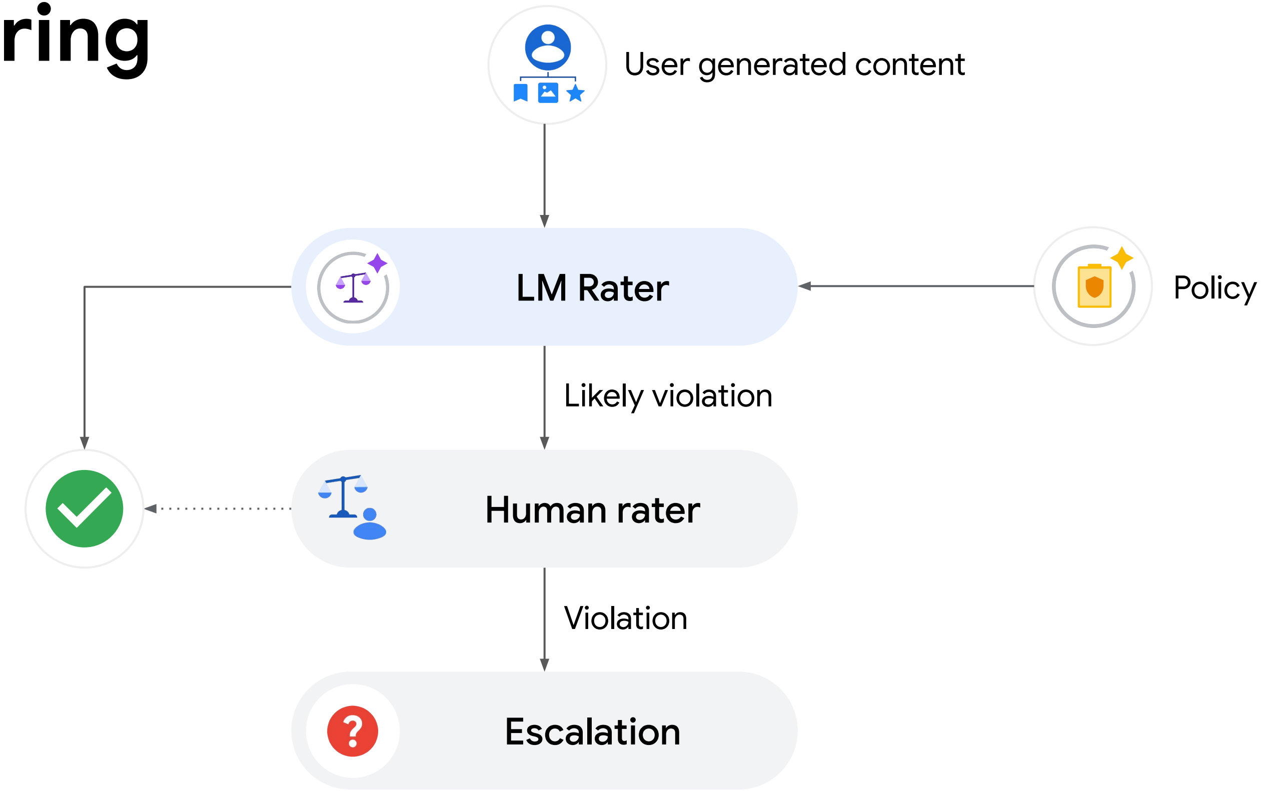


User reports

Opportunity

# Leverage large model reasoning capabilities to perform trainingless content classification

# Zero-shot pre-filtering



User generated content

LM Rater

Policy

Likely violation

Human rater

Violation

Escalation

<Misinformation and Disinformation Policy>

…

2) Comments should not make false claims that could materially discourage census participation.

3) Comments should not mislead voters about the time, place, means, or eligibility requirements of voting.

…

</Misinformation and Disinformation Policy>

Question: You are an expert content moderator. Does the following comment violate the Misinformation and Disinformation Policy?

Comment: “[COMMENT]”

Answer:

Tip: Add tags

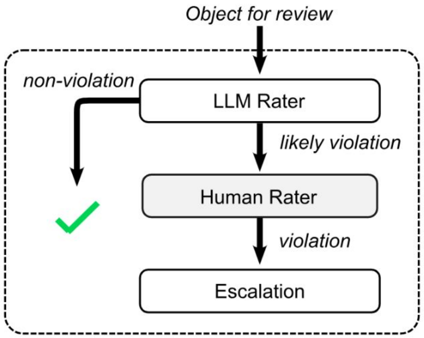
Policy

Tip: Add role

Question for the model

What to moderate

Tip: Add the beginning of the answer



Object for review

non-violation

LLM Rater

likely violation

✓

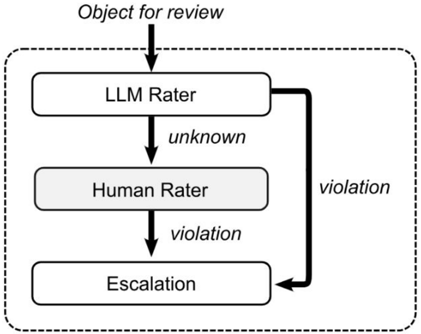
Human Rater

violation

Escalation

## Pre-filtering

Remove high-confidence non-violations from a human rater queue, focusing available resources to borderline or violative content



Object for review

LLM Rater

unknown

Human Rater

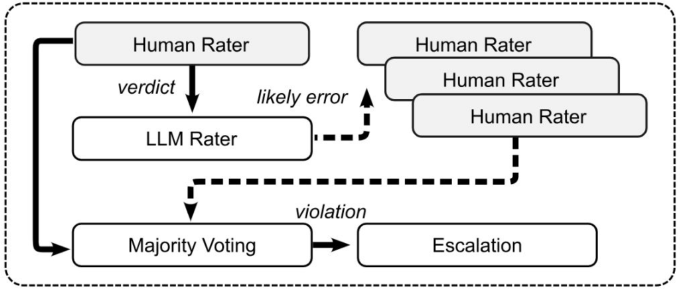
violation

violation

Escalation

## Rapid escalation

Rapidly escalate high-confidence violative content, reducing response time & toll on rater well-being



Human Rater

verdict

LLM Rater

likely error

Human Rater

Human Rater

Human Rater

violation

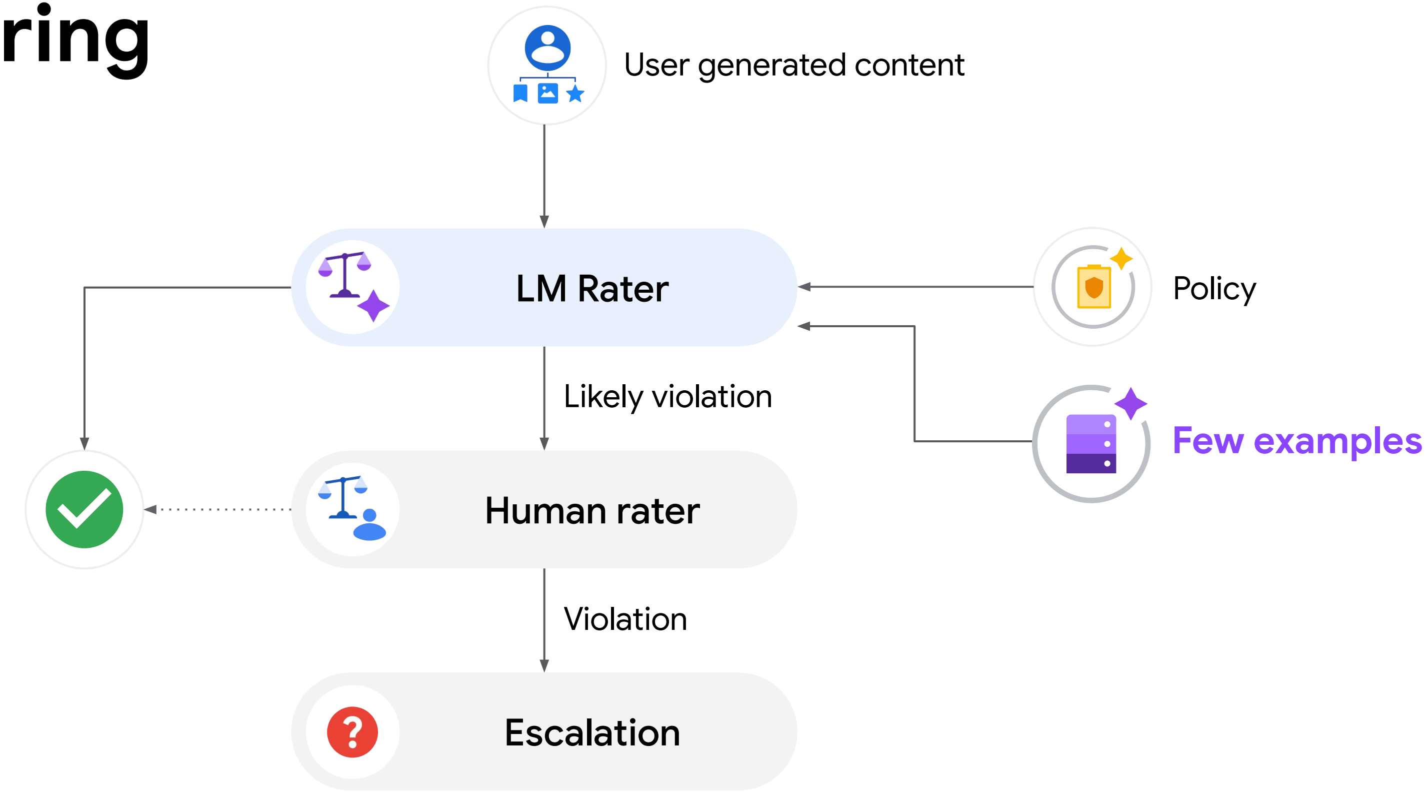
Majority Voting

Escalation

## Efficient error detection

Quickly Identify high-confidence disagreement between a human rater and an LM for error and labeling purposes. Deconflict with additional human raters

# Few-shots pre-filtering



<Misinformation and Disinformation Policy>
…
</Misinformation and Disinformation Policy>

**<examples>**
**…**
- *“No one under the age of 23 can vote” violates policy 2)*
**…**
**</examples>**

Question: **You are an expert content moderator.** Does the following comment violate the Misinformation and Disinformation Policy?

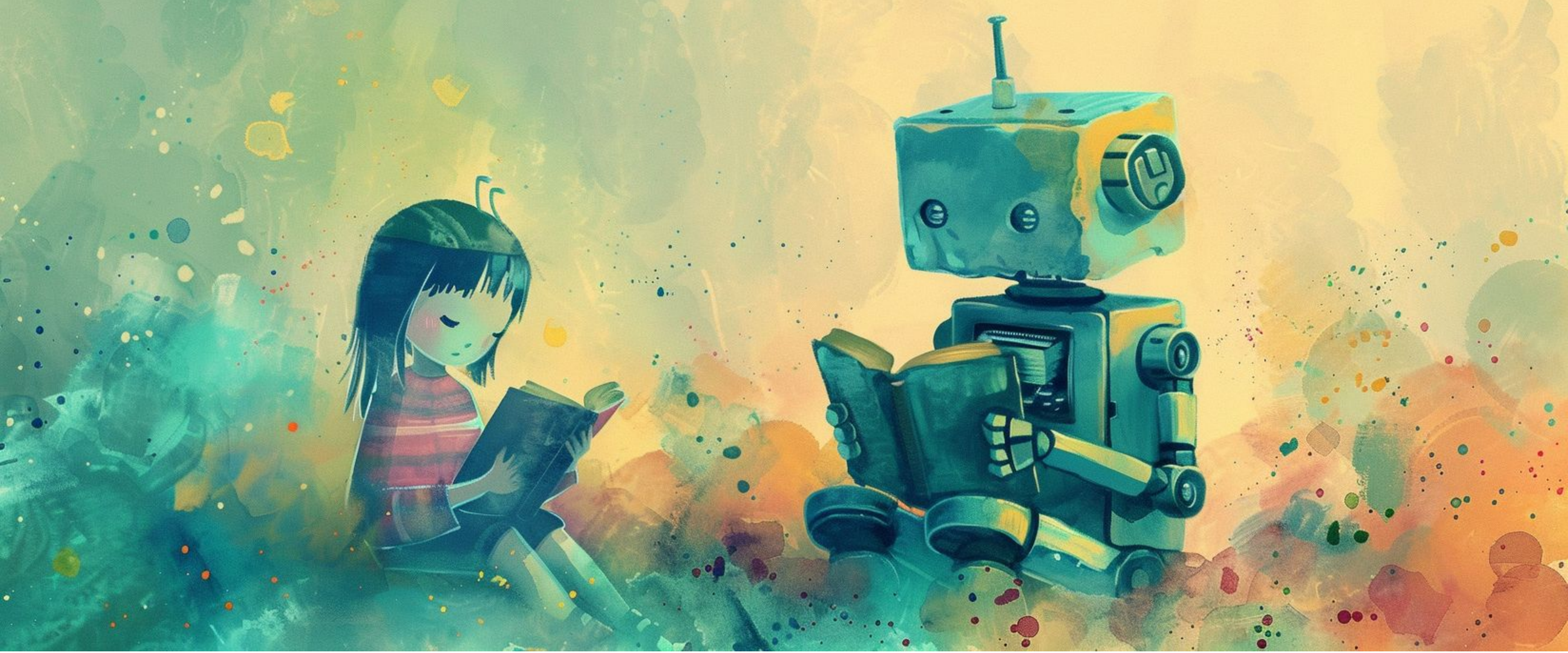
Comment: “[COMMENT]”

Answer:

## Adding examples (few shots) statically or dynamically (RAG) improves accuracy

# Experimental results

| Dataset | Static policy | Policy + RAG |
|---|---|---|
| Election Misinformation | 98.7% | 98.2% (-0.5%) |
| Hate Speech | 90.3% | 91.1% (+0.8%) |
| Violent Extremism | 89.3% | 91.1% (+1.8%) |
| Harassment | 87.2% | 90.1% (3.9%) |



Using LMs as assistant to flag key sentences helps boost human accuracy by 9–11%

Opportunity

# Leverage LM reasoning capabilities for trainingless content review scaling

Challenges

## AI understandable Policies

Policies might need to be refined
to be understandable

## Requires No-code LM integration

Ability to quickly deploy new LM fine-tuned prompts
without requiring changing services

Benefits

## Faster response to emerging threats

Reacting to new threats only requires to draft a policy
and supply a few examples

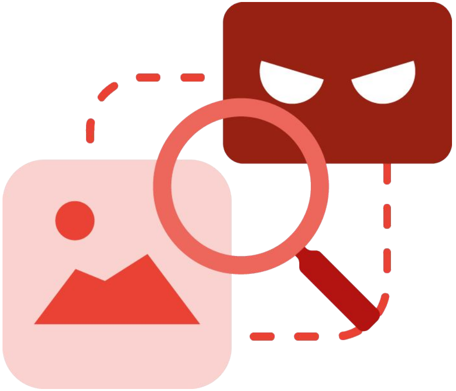
## Reduce manual toils

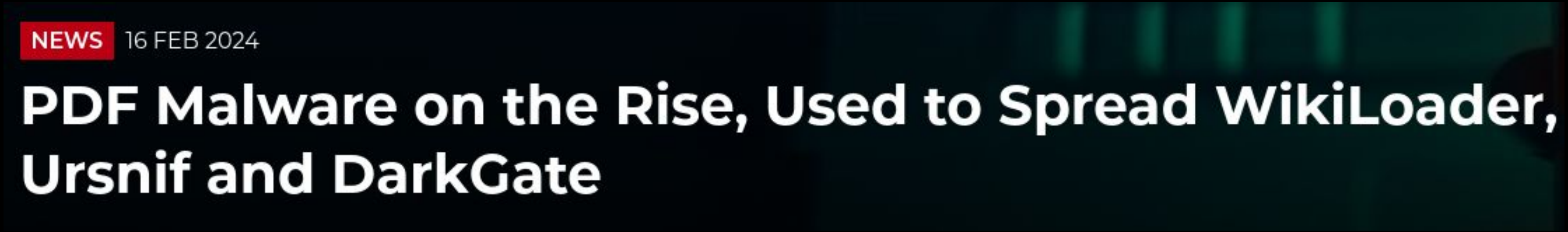
Reduce the amount of reviews done by humans

## Guardrails against review mistake

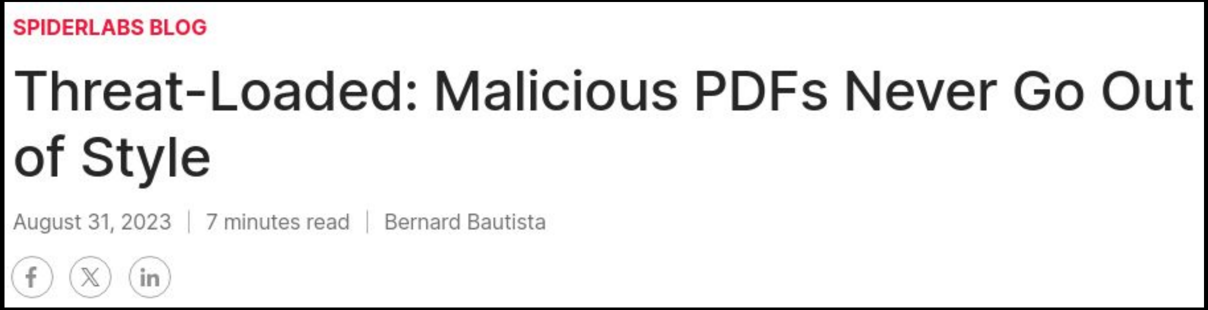
by acting as 2nd reviewer LM help spot potential
review mistakes and escalate them before they
become an issue

# Multimodal capabilities
## Analyze malicious documents



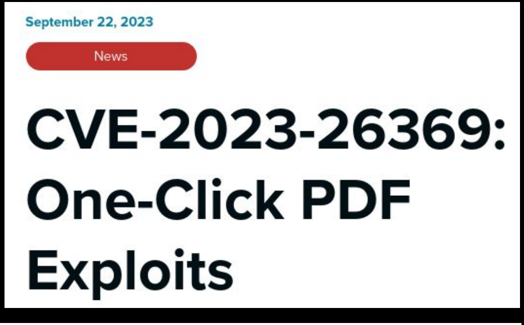

NEWS 16 FEB 2024

## PDF Malware on the Rise, Used to Spread WikiLoader, Ursnif and DarkGate


SPIDERLABS BLOG

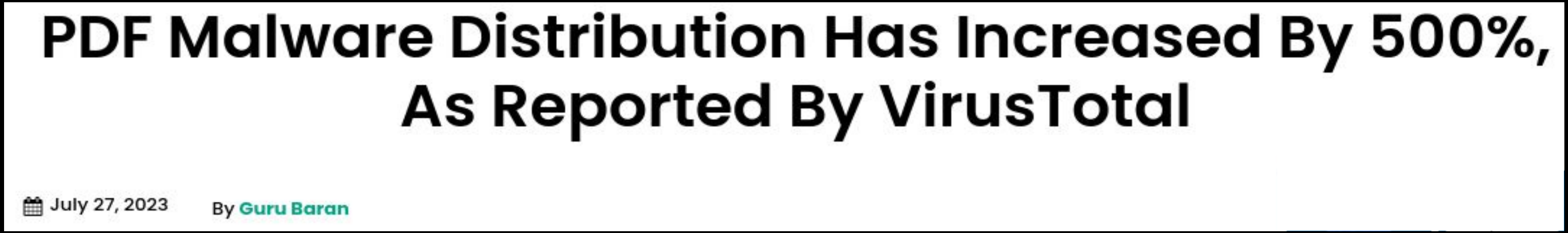
## Threat-Loaded: Malicious PDFs Never Go Out of Style

August 31, 2023 | 7 minutes read | Bernard Bautista

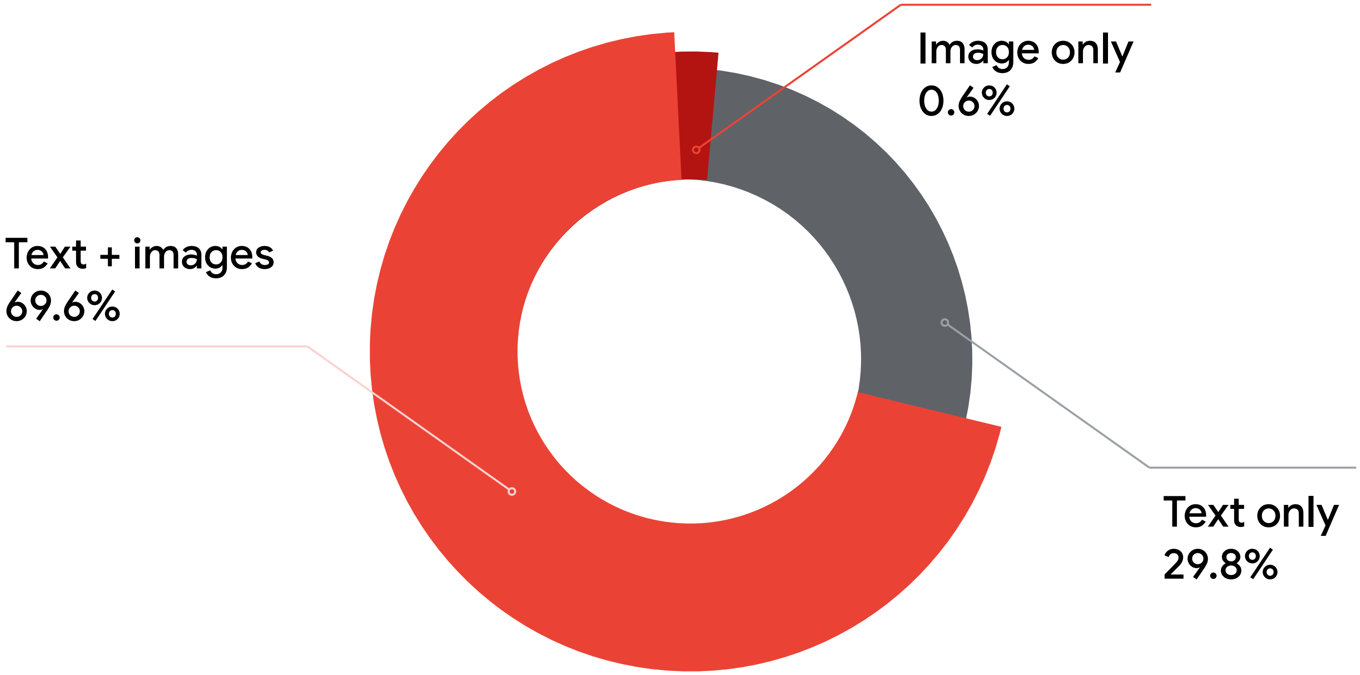

September 22, 2023

News

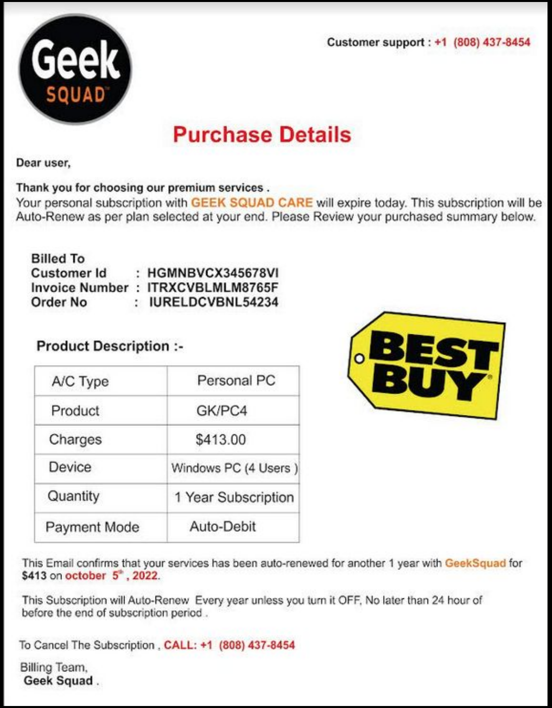
## CVE-2023-26369: One-Click PDF Exploits


## PDF Malware Distribution Has Increased By 500%, As Reported By VirusTotal

July 27, 2023 By **Guru Baran**



# Key challenge: Most malicious documents blocked by Gmail are multimodal



Customer support : +1 (808) 437-8454

Purchase Details

Dear user,

Thank you for choosing our premium services .
Your personal subscription with GEEK SQUAD CARE will expire today. This subscription will be Auto-Renew as per plan selected at your end. Please Review your purchased summary below.

Billed To
Customer Id : HGMNBVCX345678VI
Invoice Number : ITRXCVBLMLM8765F
Order No : IURELDCVBNL54234

Product Description :-

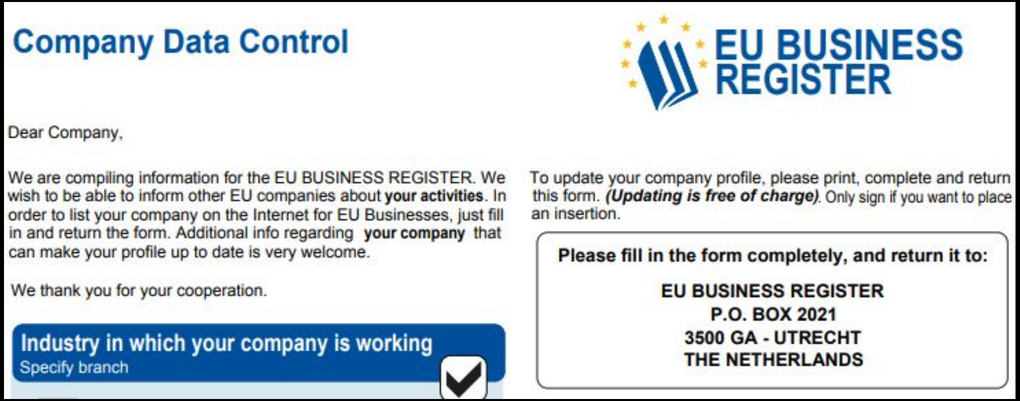
| | |
|---|---|
| A/C Type | Personal PC |
| Product | GK/PC4 |
| Charges | $413.00 |
| Device | Windows PC (4 Users ) |
| Quantity | 1 Year Subscription |
| Payment Mode | Auto-Debit |

This Email confirms that your services has been auto-renewed for another 1 year with GeekSquad for $413 on october 5th , 2022.

This Subscription will Auto-Renew Every year unless you turn it OFF, No later than 24 hour of before the end of subscription period .

To Cancel The Subscription , CALL: +1 (808) 437-8454

Billing Team,
**Geek Squad** .



Company Data Control

Dear Company,

We are compiling information for the EU BUSINESS REGISTER. We wish to be able to inform other EU companies about **your activities**. In order to list your company on the Internet for EU Businesses, just fill in and return the form. Additional info regarding **your company** that can make your profile up to date is very welcome.

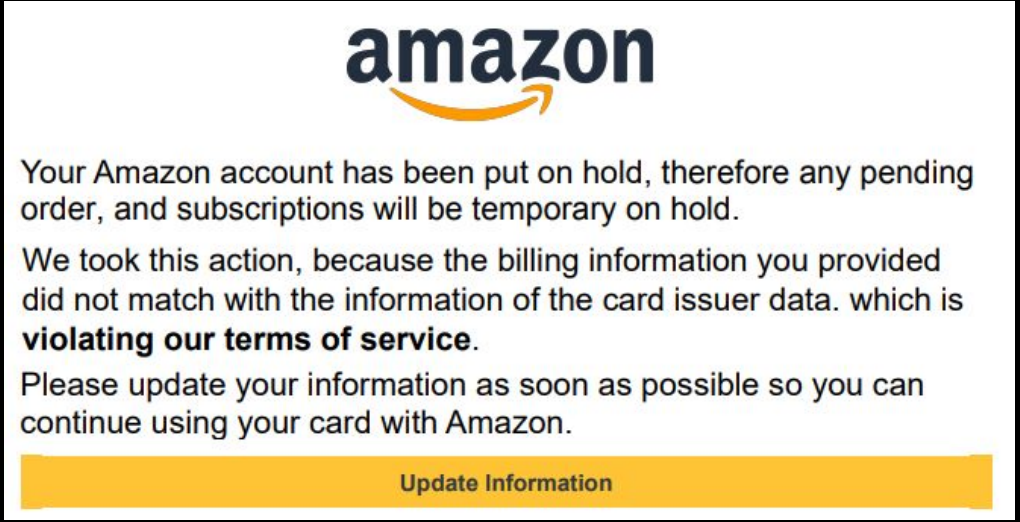
We thank you for your cooperation.

Industry in which your company is working
Specify branch ☑

To update your company profile, please print, complete and return this form. ***(Updating is free of charge)***. Only sign if you want to place an insertion.

**Please fill in the form completely, and return it to:**

**EU BUSINESS REGISTER**
**P.O. BOX 2021**
**3500 GA - UTRECHT**
**THE NETHERLANDS**



Your Amazon account has been put on hold, therefore any pending order, and subscriptions will be temporary on hold.

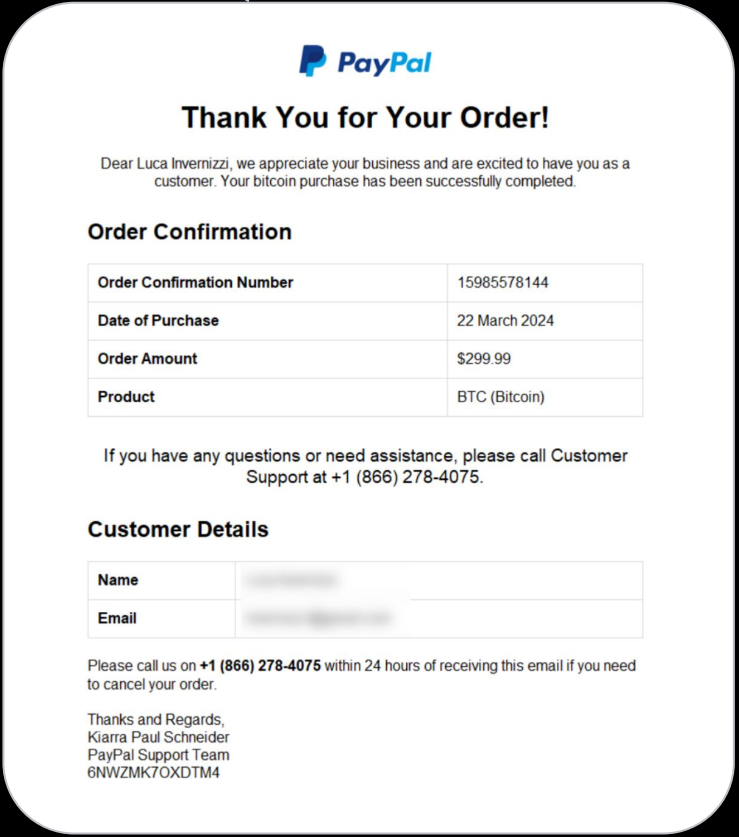
We took this action, because the billing information you provided did not match with the information of the card issuer data. which is **violating our terms of service**.

Please update your information as soon as possible so you can continue using your card with Amazon.

Using only prompt-tuning Gemini Pro achieves 91% accuracy on Virus Total dataset

However our specialized model achieve 99% accuracy while being ~100x faster to run

# Leverage large model multimedia capabilities to analyze multimodal malicious documents



PayPal

## Thank You for Your Order!

Dear Luca Invernizzi, we appreciate your business and are excited to have you as a customer. Your bitcoin purchase has been successfully completed.

### Order Confirmation

| | |
|---|---|
| **Order Confirmation Number** | 15985578144 |
| **Date of Purchase** | 22 March 2024 |
| **Order Amount** | $299.99 |
| **Product** | BTC (Bitcoin) |

If you have any questions or need assistance, please call Customer Support at +1 (866) 278-4075.

### Customer Details

| | |
|---|---|
| **Name** | [redacted] |
| **Email** | [redacted] |

Please call us on **+1 (866) 278-4075** within 24 hours of receiving this email if you need to cancel your order.

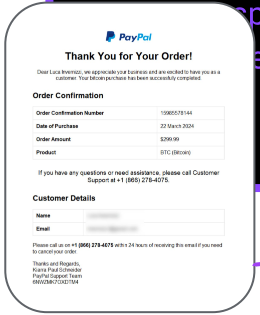
Thanks and Regards,
Kiarra Paul Schneider
PayPal Support Team
6NWZMK7OXDTM4

Opportunity

# Leverage LM multimodal to offer meaningful explanation

This document is likely a phishing attempt impersonating Paypal and should not be trusted. Here are some reasons why:

- Suspicious phone number: The phone number "+1 (866) 278-4075" does not match the official PayPal customer support phone numbers.
- Urgency or pressure tactics: The document urges you to call within 24 hours to cancel the order, creating a sense of urgency and pressuring you to act quickly.
- Call back requested: The document asks you to call a specific phone number, which could be used to steal your ersonal information.



PayPal

**Thank You for Your Order!**

Dear Luca Invernizzi, we appreciate your business and are excited to have you as a customer. Your bitcoin purchase has been successfully completed.

**Order Confirmation**

| | |
|---|---|
| **Order Confirmation Number** | 15985578144 |
| **Date of Purchase** | 22 March 2024 |
| **Order Amount** | $299.99 |
| **Product** | BTC (Bitcoin) |

If you have any questions or need assistance, please call Customer Support at +1 (866) 278-4075.

**Customer Details**

| | |
|---|---|
| **Name** | [redacted] |
| **Email** | [redacted] |

Please call us on **+1 (866) 278-4075** within 24 hours of receiving this email if you need to cancel your order.

Thanks and Regards,
Kiarra Paul Schneider
PayPal Support Team
6NWZMK7OXDTM4

Perfectly understood the image complex content

Identified key discrepancy from real invoice

Retrieval data will be key here to get correct number

Correctly assessed risk and consequence

Opportunity

# Leverage LM multimodal capabilities to detect multimodal malicious documents

Challenges

## Fine-tuning required

Getting the best performance out of LM requires full fine-tuning rather than prompt engineering

## Prohibitively hard to scale

LM computation cost makes using LM at Gmail scale infeasible but great for small scale. Large scale requires a specialized model

Benefits

## Deal with multimodal attacks

LM are able to jointly process images, text, code, giving them an edge when understanding multimodal threats

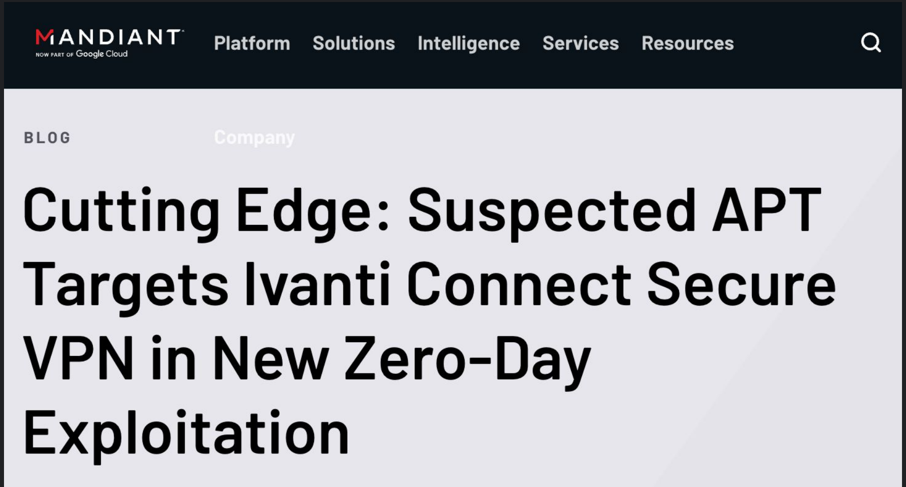
## Generalize across formats

Semantic understanding of the threats allows the detection to be filetype and metadata agnostic

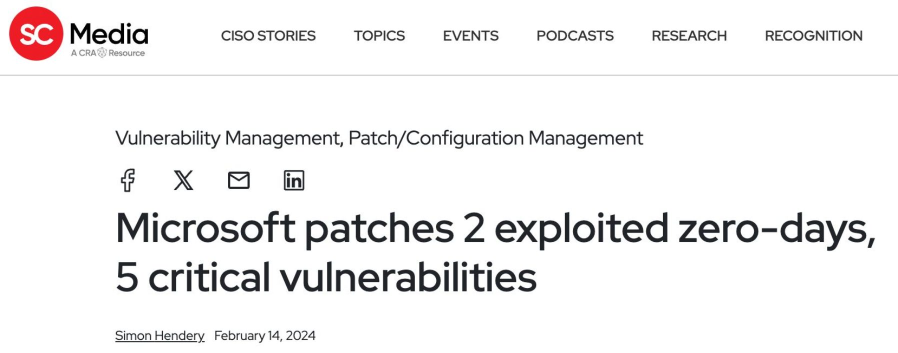
## Act as an analyst

LM answers go beyond classification: providing analyst-level capabilities that are easier for users to understand

# Coding capabilities
## Secure codebase



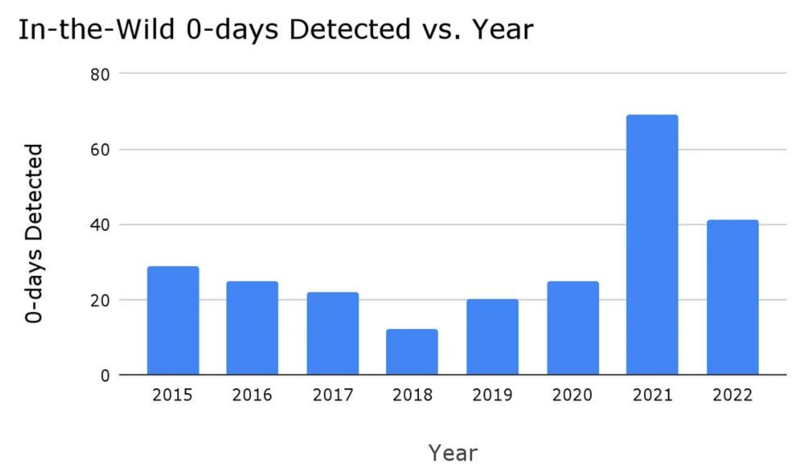
# Cutting Edge: Suspected APT Targets Ivanti Connect Secure VPN in New Zero-Day Exploitation



Vulnerability Management, Patch/Configuration Management

# Microsoft patches 2 exploited zero-days, 5 critical vulnerabilities

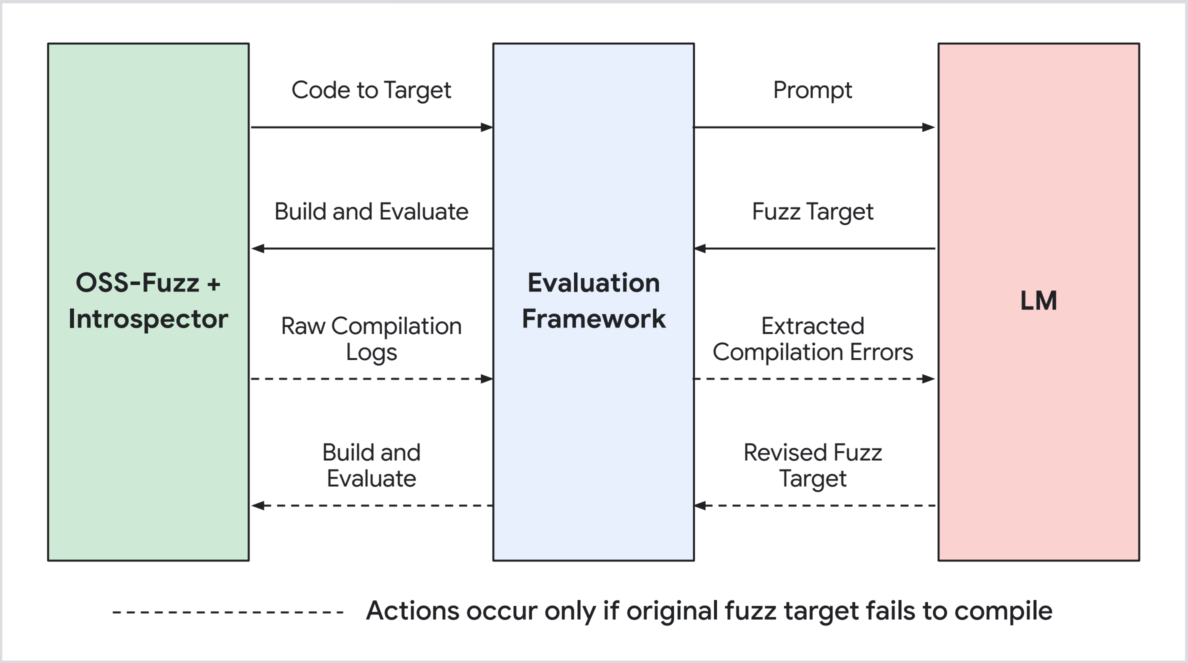
Simon Hendery February 14, 2024



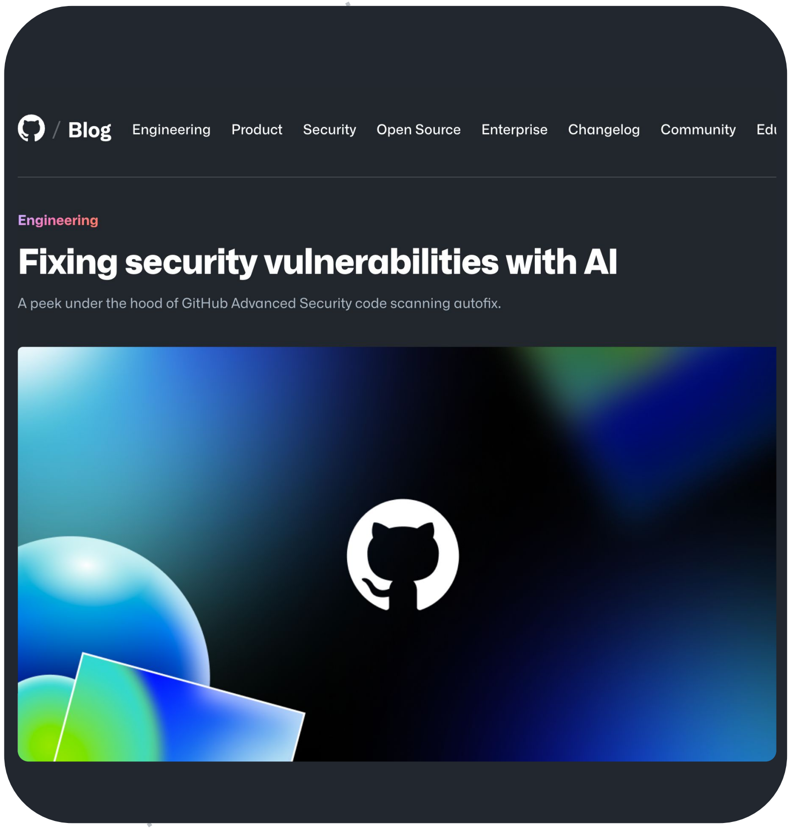
https://security.googleblog.com/2023/07/the-ups-and-downs-of-0-days-year-in.html

✓ Early success

# LM code understanding can be used to enhance fuzzers harness



# GitHub is actively developing an assistant to help detect and fix vulnerabilities

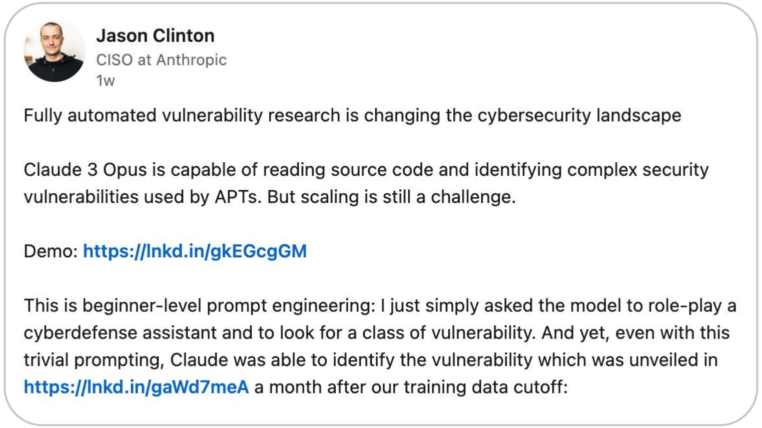


Engineering

Fixing security vulnerabilities with AI

A peek under the hood of GitHub Advanced Security code scanning autofix.

https://github.blog/2024-02-14-fixing-security-vulnerabilities-with-ai/



**Jason Clinton**
CISO at Anthropic
1w

Fully automated vulnerability research is changing the cybersecurity landscape

Claude 3 Opus is capable of reading source code and identifying complex security vulnerabilities used by APTs. But scaling is still a challenge.

Demo: https://lnkd.in/gkEGcgGM

This is beginner-level prompt engineering: I just simply asked the model to role-play a cyberdefense assistant and to look for a class of vulnerability. And yet, even with this trivial prompting, Claude was able to identify the vulnerability which was unveiled in https://lnkd.in/gaWd7meA a month after our training data cutoff:

Opportunity

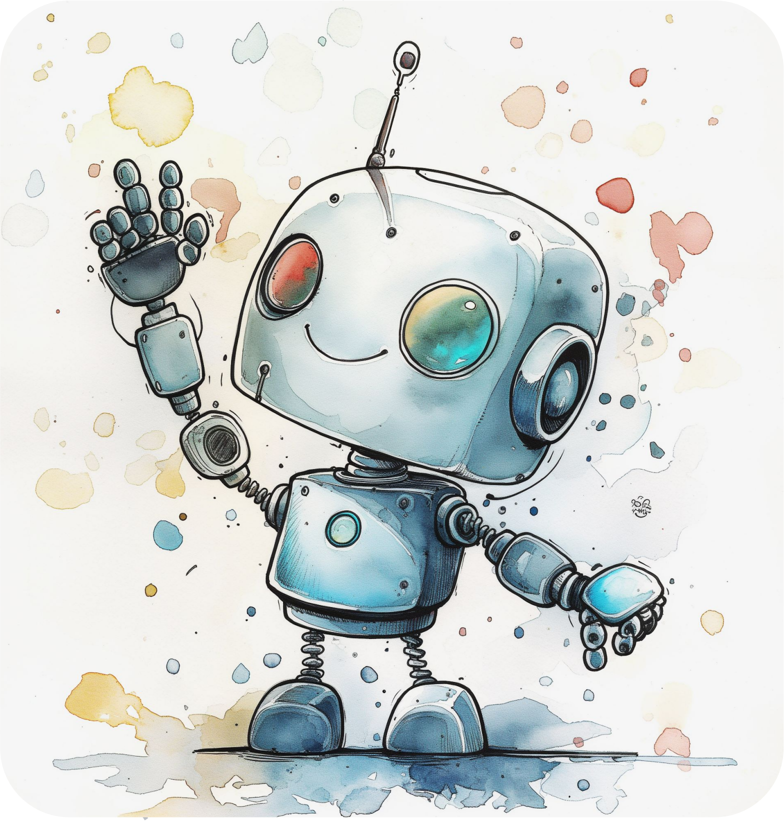
# Leverage large model code understanding to find and patch code vulnerabilities

# Hype alert
# New vulnerability detection benchmark shows current results don't generalize

https://arxiv.org/abs/2403.18624

✓ Early success

# A Code LM was able to successfully patch 15% of the simple vulnerabilities found by sanitizers

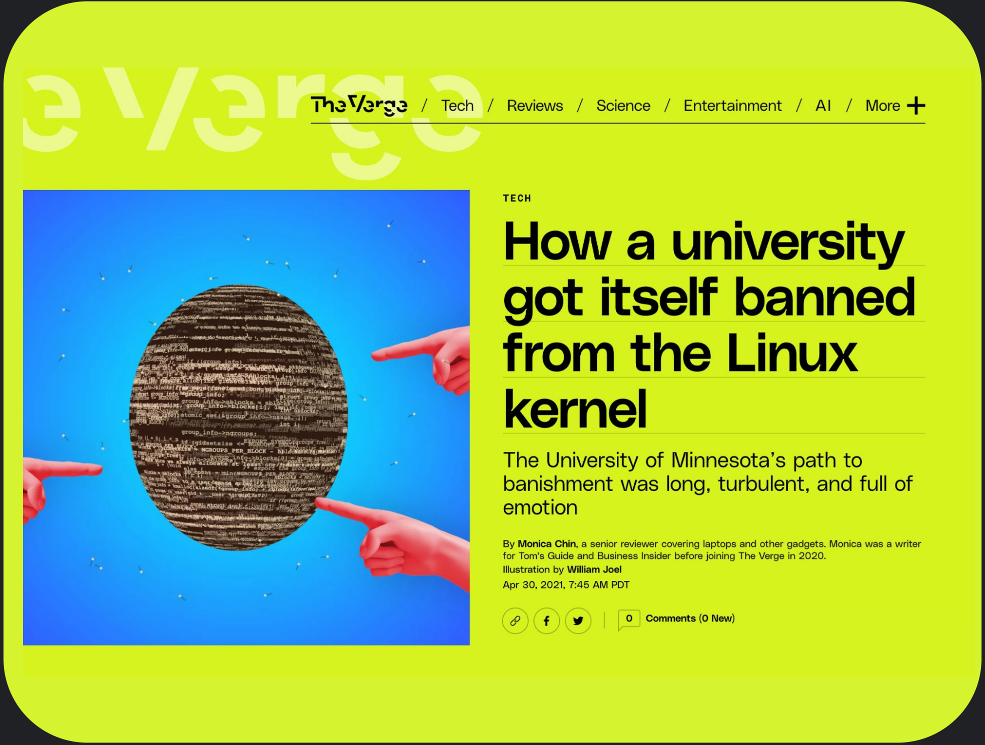


https://research.google/pubs/ai-powered-patching-the-future-of-automated-vulnerability-fixes/

# Some interesting behaviors

- ✓ AI can patch code
- ✓ Some bug are easier than other
- ✓ Able to add a mutex to fix a race condition
- ✓ Can fix data leak by removing pointer uses

- ✗ “Only” 15% success rate long way to go
- ✗ Commented out code to solve the problem…
- ✗ Rewrote the code to run sequentially
- ✗ Deleted unit tests causing the detection



TECH

How a university got itself banned from the Linux kernel

The University of Minnesota's path to banishment was long, turbulent, and full of emotion

By **Monica Chin**, a senior reviewer covering laptops and other gadgets. Monica was a writer for Tom's Guide and Business Insider before joining The Verge in 2020.
Illustration by **William Joel**
Apr 30, 2021, 7:45 AM PDT

# Model patching accuracy in nowhere near the level needed for production

Opportunity

# Leverage lm coding capabilities to find and patch code vulnerabilities

Challenges

### Validation is very complex

Validating that a patch fixes a vulnerability without breaking anything requires extensive tests and/or manual review

### Dataset creation

Creating the right dataset is difficult: requires a large scale manual effort by experts

### Complex training

Getting the best performance requires a complex interplay of training techniques and a lot of compute

Benefits

### Help find vulnerabilities faster

Complement to fuzzing assuming precision is good enough to not create too many false positives

### Eliminate windows of vulnerability

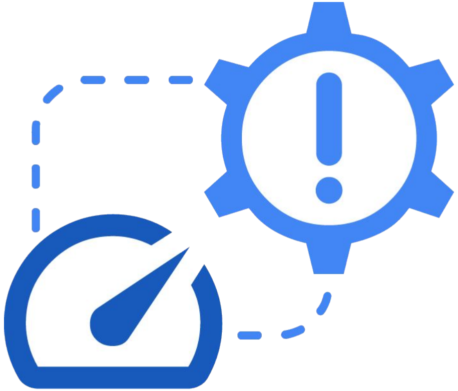
Vuln detection + patch generation if accurate has the potential to eliminate the vulnerability windows by offering a fix at commit time

### Reduce manual burden

Help triage bug reports and generate fixes

# Generative capabilities

## Speed-up incident response



# During incident response speed is of the essence

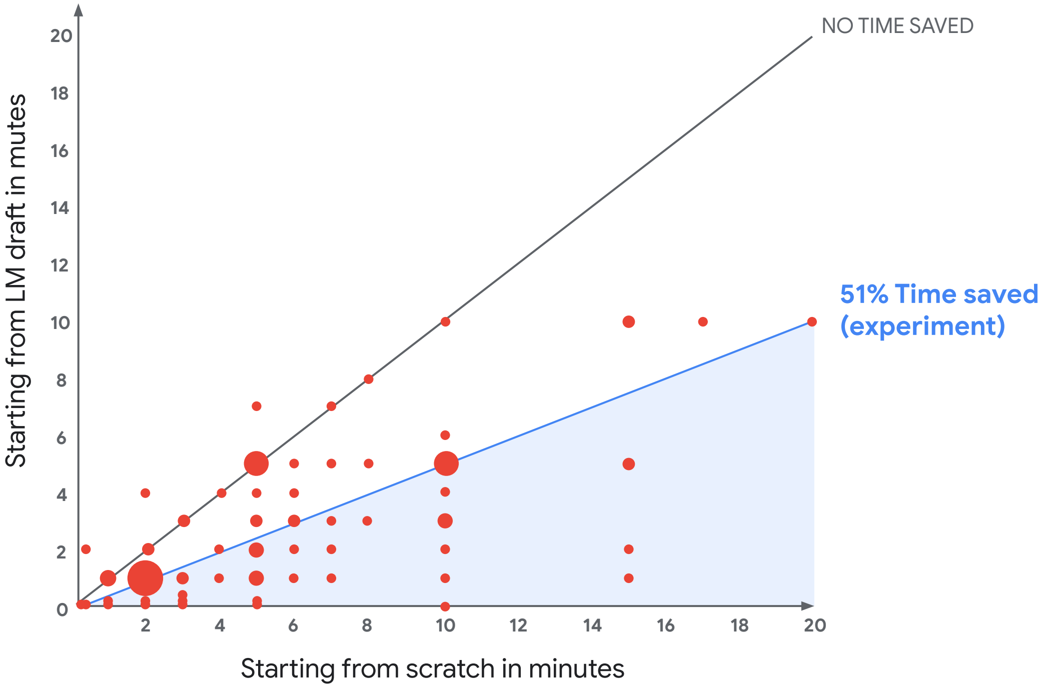
Opportunity

# Leverage large model generative capabilities to speed-up incident response

✓ Early success

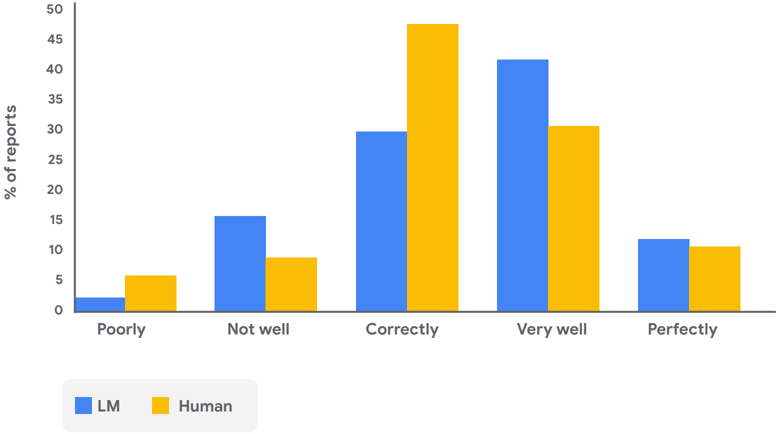
# LM are able to help incident teams write incident summaries 51% faster

Time spent (in minutes) writing an incident summary

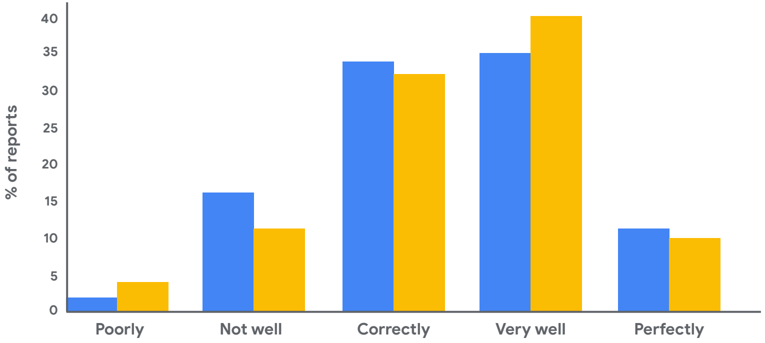


https://security.googleblog.com/2024/04/accelerating-incident-response-using.html

How well does this summary follow the **writing guidelines?**



How well does this summary cover the incident's **key points?**



# LMs are comparable to humans when writing incident summaries

```
<Security Incident>
<Title> [tool_name_verdict] Abuse verdict for project id: xyz.</Title>
<Metadata> This ticket was filled and submitted on the 2023-10-01. It was marked with the labels:
“Investigation” and “AB”.</Metadata>
<Description> Counter-Abuse has issued an abuse verdict against a GCP project.</Description>
<Additional Information> The incident was reported through the xyz pipeline with a policy violation
of “COIN_MINING”.
The infraction can be found in the project xyz.</Additional Information>
<Date Incident> 2023-10-01 11:50:19</Date Incident><Incident Causes> The identified causes are:
MISCONFIGURATION, WEAK_OR_NO_PASSWORD</Incident Causes><Actions Taken> The following actions were
taken:
1) Action1
2) Action2</Actions Taken>
<Software Involved> Software1</Software Involved>
<Sensitive Data> - NONE, TEST</Sensitive Data>
<Mitigation History><Comment index=”1” author=”user1@domain.com”> Looks like there was a CPU spike:
URL around 05:00. Running application1 now.</Comment>
<Comment index=”2” author=”user3@domain.com”> Instance compromised, shutting it down</Comment>
<Comment index=”3” author=”user4@domain.com”> InstanceMetadata</Comment>
<Comment index=”4” author=”user@domain.com”> Get additional information on InstanceMetadata:
URL`<Code Section/>`</Comment>
<Comment index=”5” author=”user3@domain.com”> Looks like it was compromised through  successfully
authentication as root account using SSH with password authentication: `<Code Section/>`</Comment>
<Comment index=”6” author=”user3@domain.com”> A malicious cron job was created on the machine
`<Code Section/>`. The cron job downloaded a bash script from IP and executed it. The script was
not present under `<Code Section/>` at the time of the investigation `<Code Section/>`</Comment>
<Comment index=”10” author=”user3@domain.com”> Exec update sent.</Comment>
</Mitigation History>
```

# Getting good results requires very well structured data and prompts

Opportunity

# Leverage LM generative capabilities to speed-up incident response

Challenges

### Complex data input

Incident data must be very well structured to get good results

### Only speed up summarization

So far LM are only able to help with summarization not doing the root cause analysis

### Requires human in the loop

Summaries must be proofread by analysts to ensure correctness and completeness

Benefits

### 51% faster summarization

LM helps reduce incident time by making the summarization 2x faster

### More consistent summaries

LM are more consistent at following guidelines than humans, leading to more consistent summaries overall

# Capabilities recap

**Reasoning capabilities**
Scale content review

**Multimodal understanding**
Analyze malicious documents

**Code understanding**
Secure code

**Generative capabilities**
Speed up incident response

# Takeaways

AI will give the advantage back to the defenders

AI is also driving advanced offensive capabilities and lowering the technical bar

More research is urgently needed to harness AI cybersecurity capabilities

## Apply Today

Review your current defenses to identify which would benefit the most from AI

## In the next 6 months

Increase defense in depth by adding at least one AI powered defense

Increase preparedness by educating your workforce about the rise of AI offensive capabilities

# Thank you

Scan me with your phone

**Presentation slides and recording available here**

https://elie.net/lmsec24