

AI for Cybersecurity

Get Started Today



**Elie
Bursztein**
Google Deepmind

with the help of **many** Googlers and external collaborators





Presentation slides
<https://elie.net/ai24>



AI will
revolutionize
cybersecurity

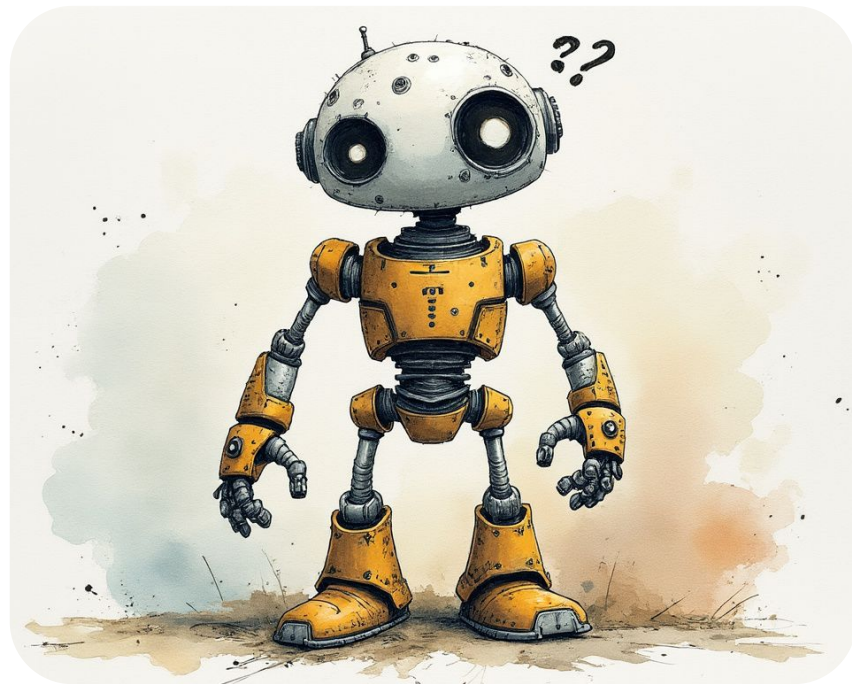




Lot of hype and buzz around AI & Cybersecurity

Beyond the buzz and hype around AI

Where do I start?





AI is disrupting the cybersecurity balance



Bad actors



Lowering
the bar



Defenders



Scaling
capacity



How AI is actively
enhancing offensive
capabilities

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



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

FORBES > INNOVATION

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

 INDEPENDENT

Push notifications




NEWS SPORTS VOICES CULTURE LIFESTYLE TRAVEL PREMIUM

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 Africa Americas Asia Australia China More

Watch



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

Multimodal Deep-fake
offensive capabilities is
current the key risk to
get ready for



**Where should I start using
AI to scale my defenses?**

Current AI capabilities assessment



Reasoning

Readiness ♦♦

Use LM to reason about complex questions such as deciding if it is an intrusion, doing root cause analysis or manual reviews



Multimodal

Readiness ♦♦♦

Understanding complex documents including video and images



Generation

Readiness ♦

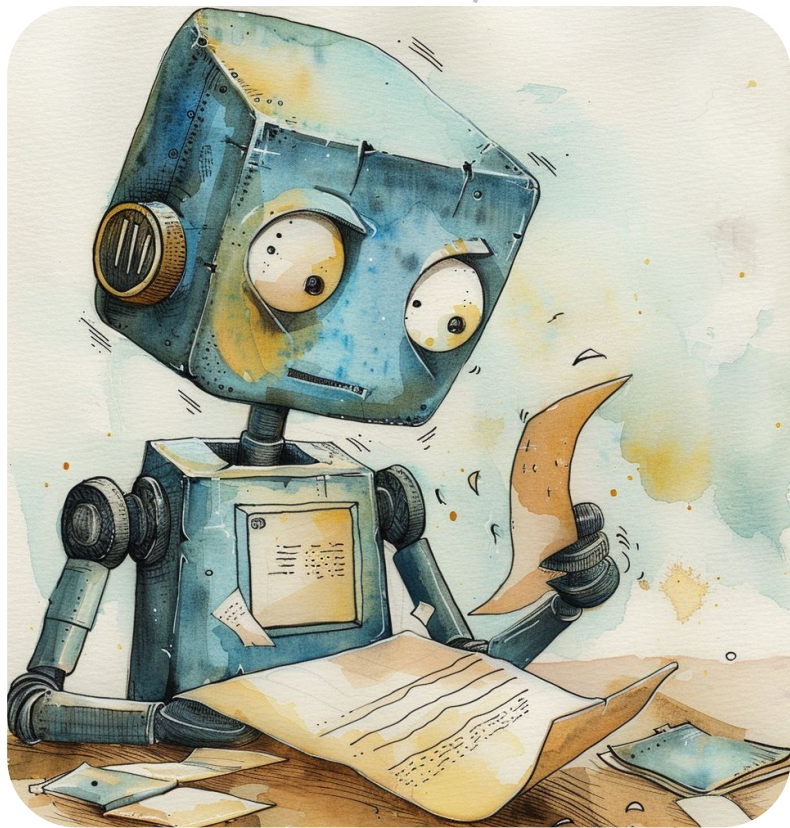
Generate complex documents and artifacts configuration, code patch, firewall, malware



Synthetisation

Readiness ♦♦♦♦

Retrieve data and provide a summary of what the content is to help humans - e.g incident response



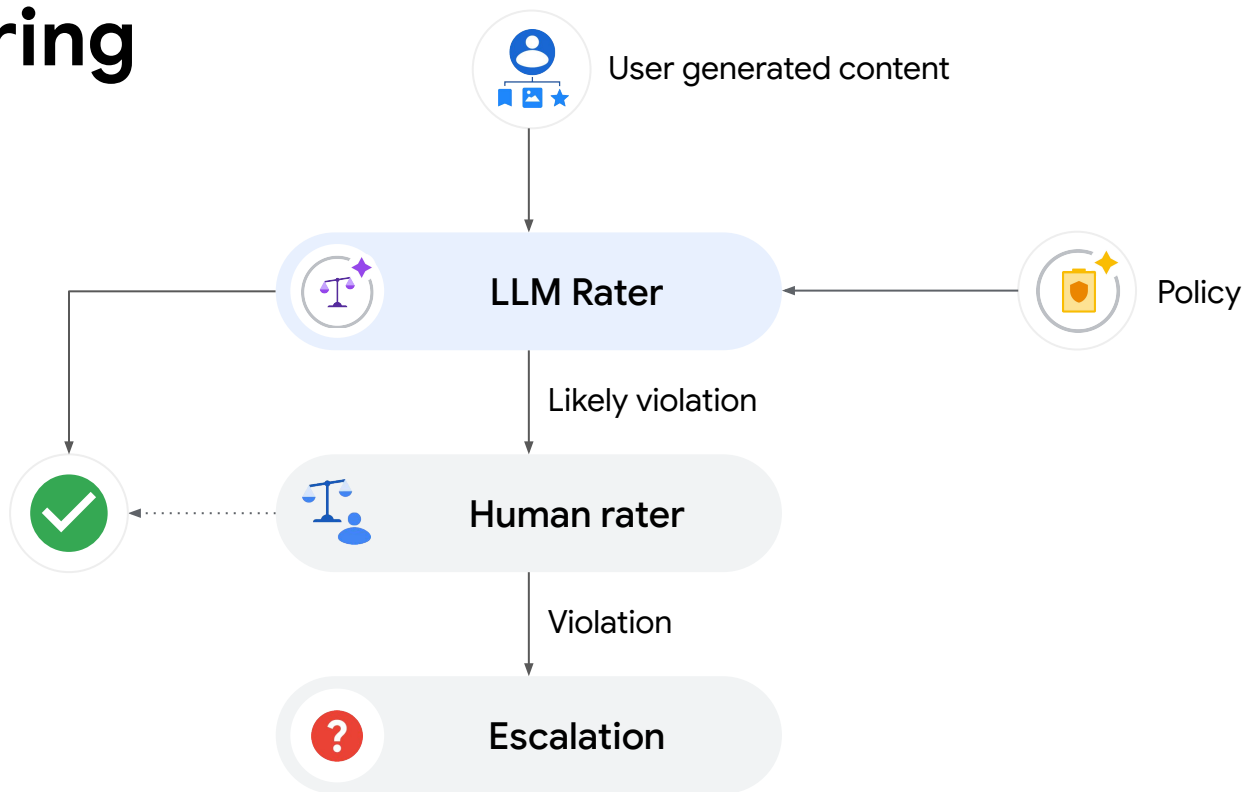
Opportunity

Leverage large model reasoning & synthetisation capabilities to perform trainingless content classification and summarization



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

Zero-shot pre-filtering



<Misinformation and Disinformation Policy>

Tip: Add tags

...

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.

Policy

...

</Misinformation and Disinformation Policy>

Tip: Add role

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

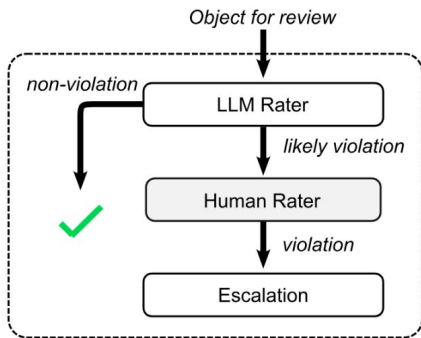
Question for the model

Comment: "[COMMENT]"

What to moderate

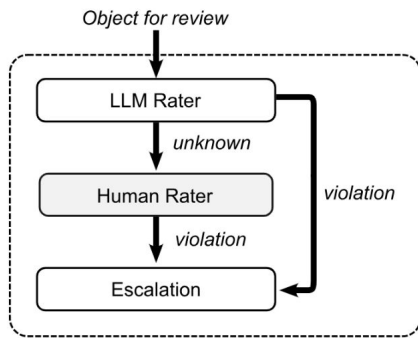
Answer:

Tip: Add the beginning of the answer



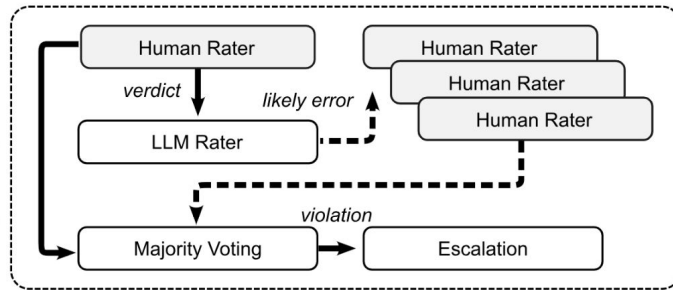
🚫 Pre-filtering

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



📈 Rapid escalation

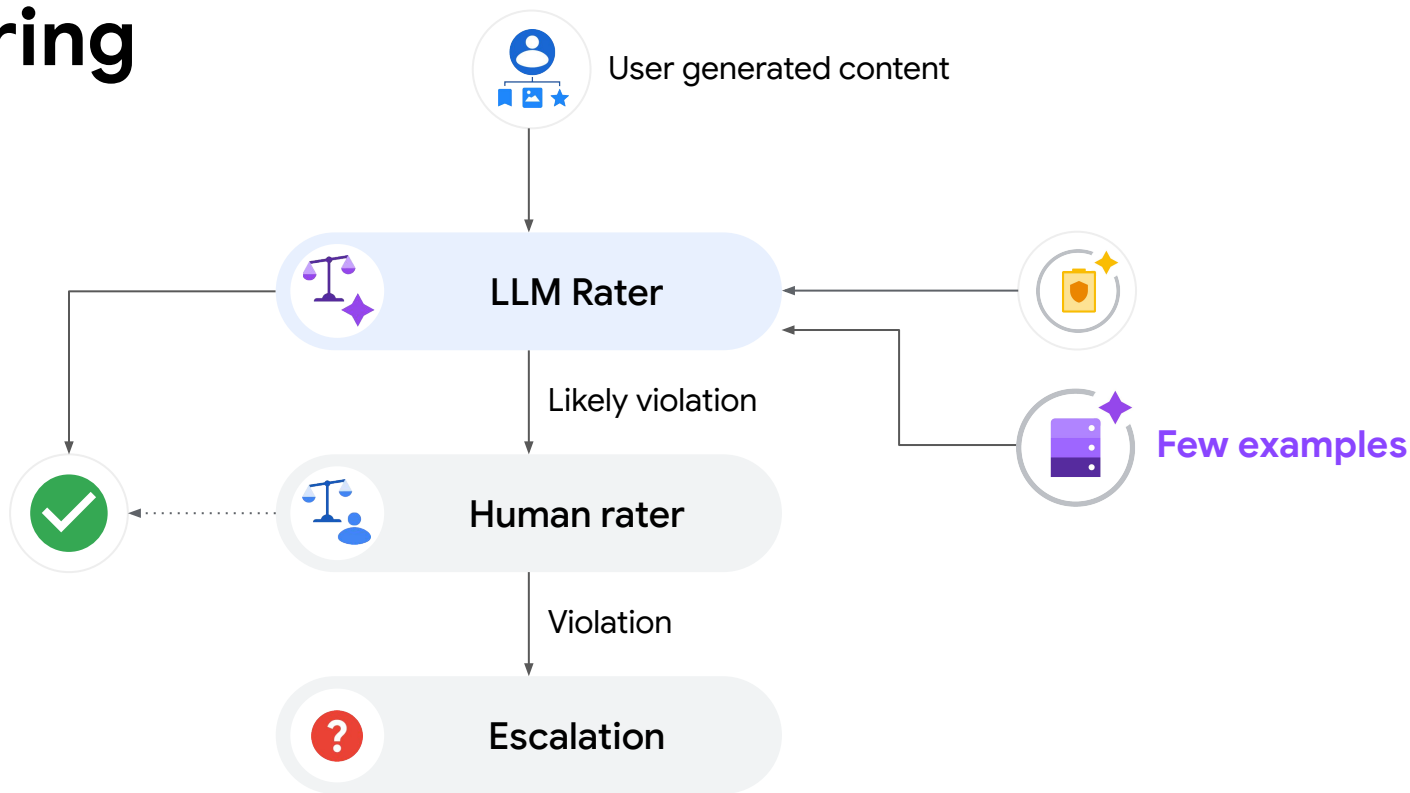
Rapidly escalate high-confidence violative content, reducing response time & toll on rater well-being



⚠️ Efficient error detection

Quickly Identify high-confidence disagreement between a human rater and an LLM 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?

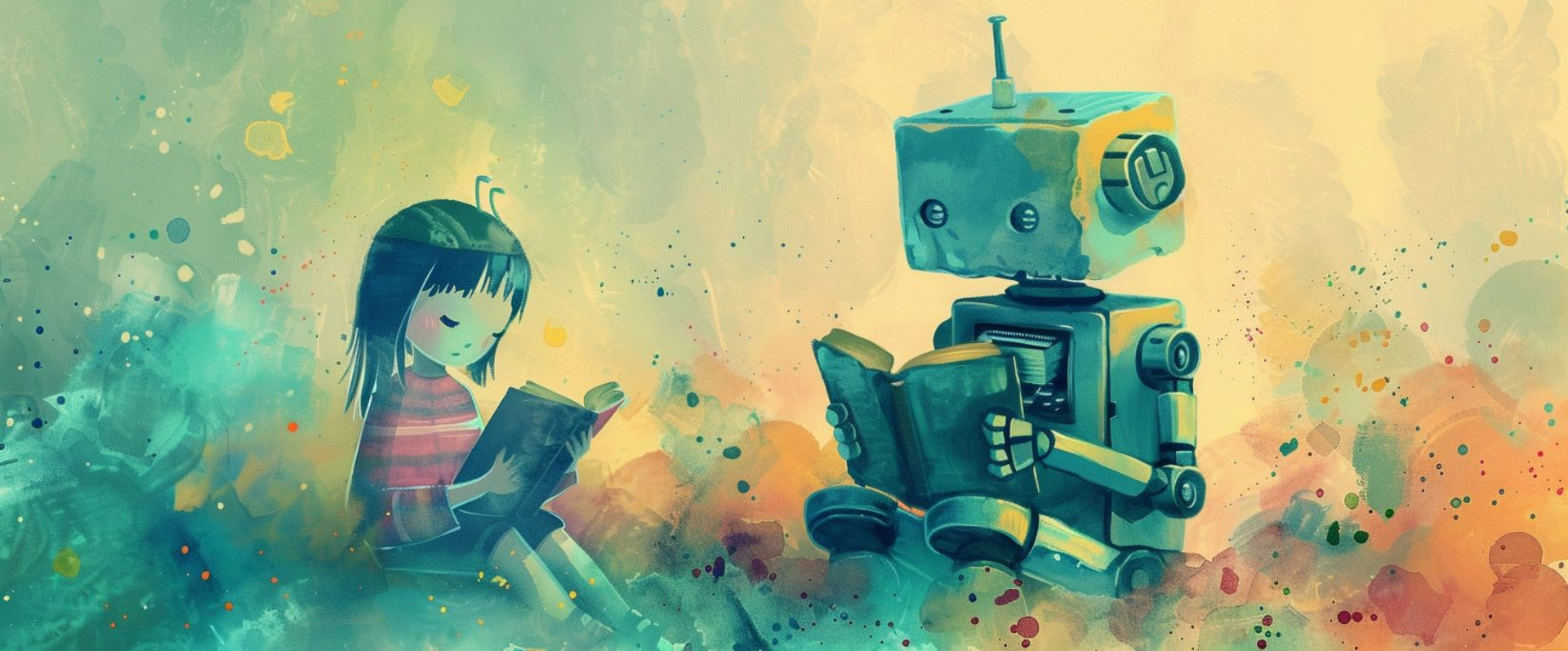
Comment: "[COMMENT]"

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



Opportunity

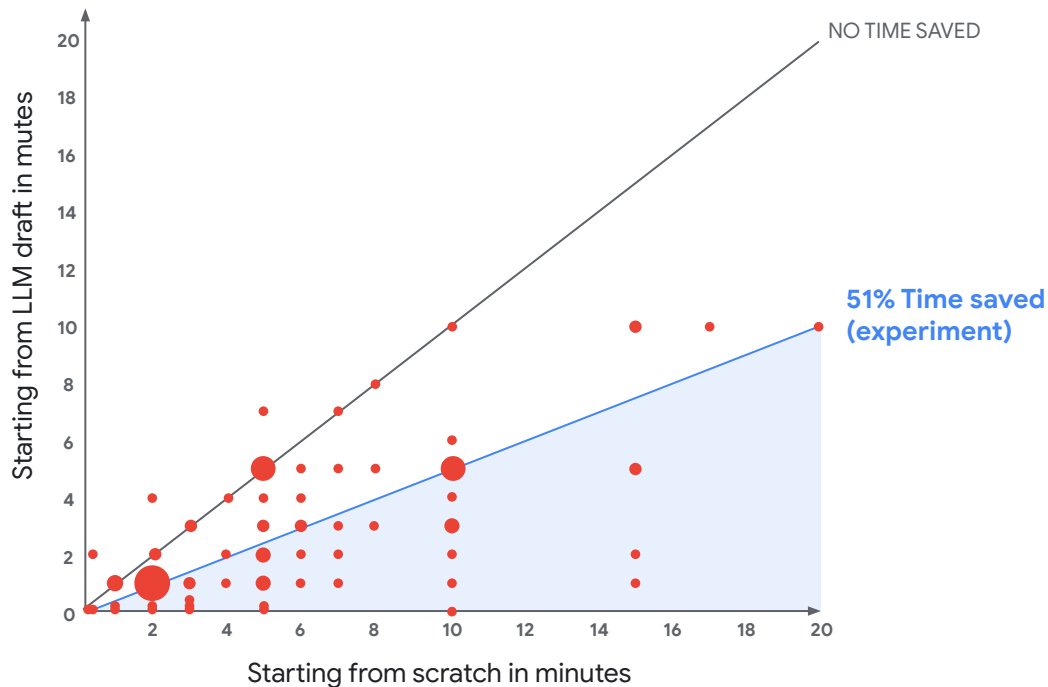
**Leverage large model
generative capabilities
to speed-up incident
response**



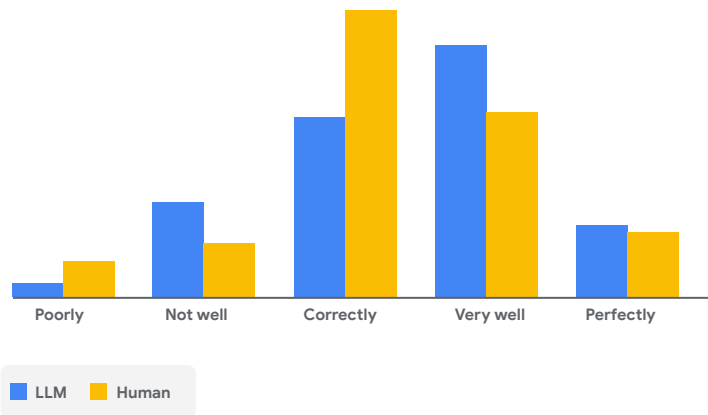
Early success

LLM are able to help
incident teams **write**
incident summaries
51% faster

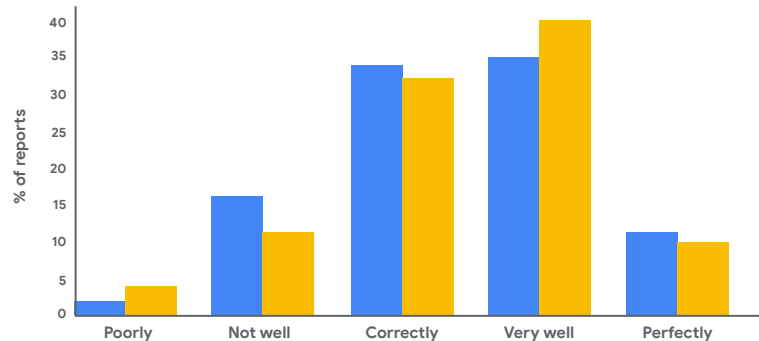
Time spent (in minutes) writing an incident summary



How well does this summary follow the writing guidelines?



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



LLMs are comparable to humans when writing incident summaries

Takeaways



Multimodal Deep-fake is top concern



AI powered cybersecurity capabilities
are coming online



Significant research is still need to get
to the most complex use-cases



AI will ultimately tip the balance in favor of the defenders



Thank You

<https://elie.net/ai24>

