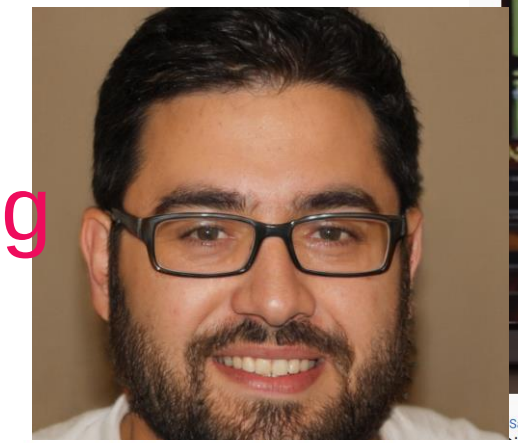
Welcome AmplifyIT Oshkosh!
Generative AI
Risks, Rewards, Tradeoffs

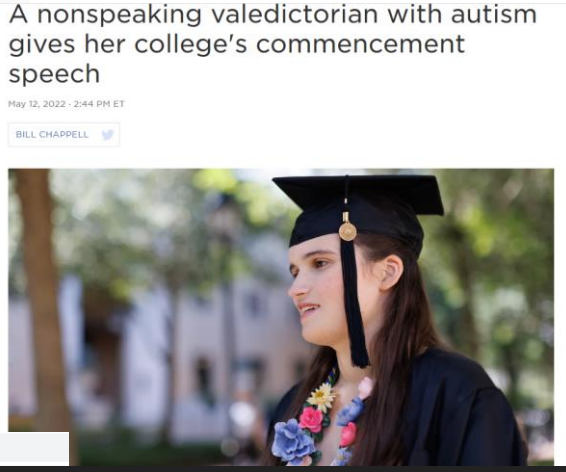
Michele Boland Architect and Evangelist
Check Point Office of the CTO
linkedin.com/in/micheleboland
+1 972-824-1880
9 APR 2024

YOU DESERVE THE BEST SECURITY

Deepfake technologies, Gen AI risks, rewards, tradeoffs

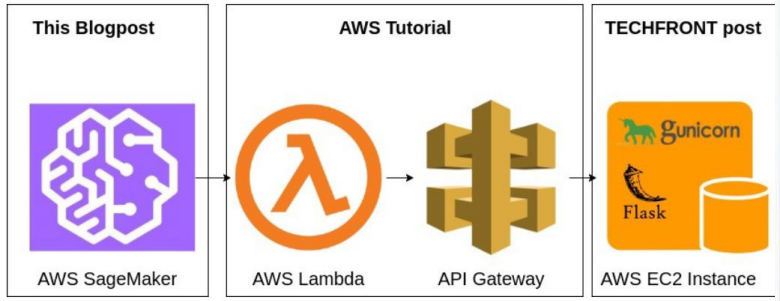
Deepfakes and social engineering bots



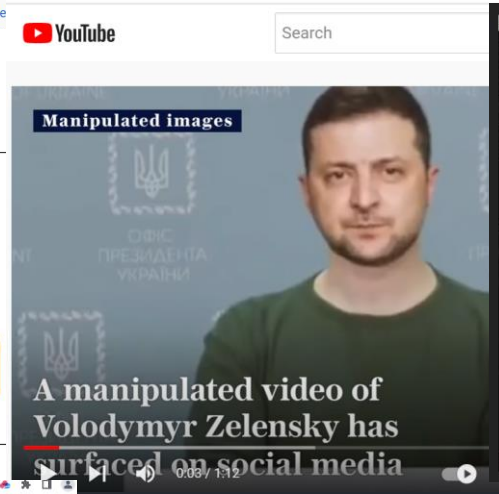


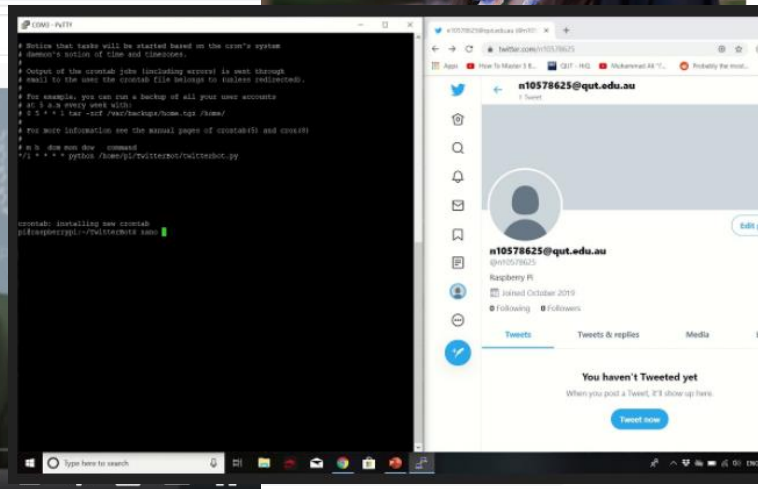
anfreeman
Morgan Freeman - A Deepfake Singularity

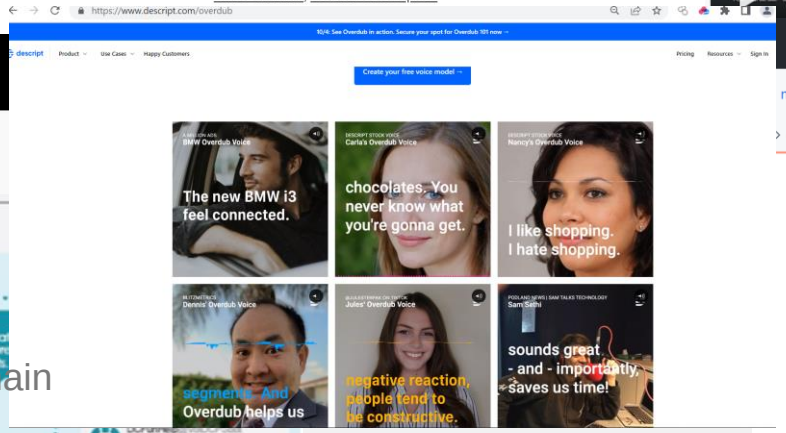
In this blogpost, we will cover the first task in detail. Two others are covered in AWS Tutorial, TECHFRONT post. The final architecture will look like this:

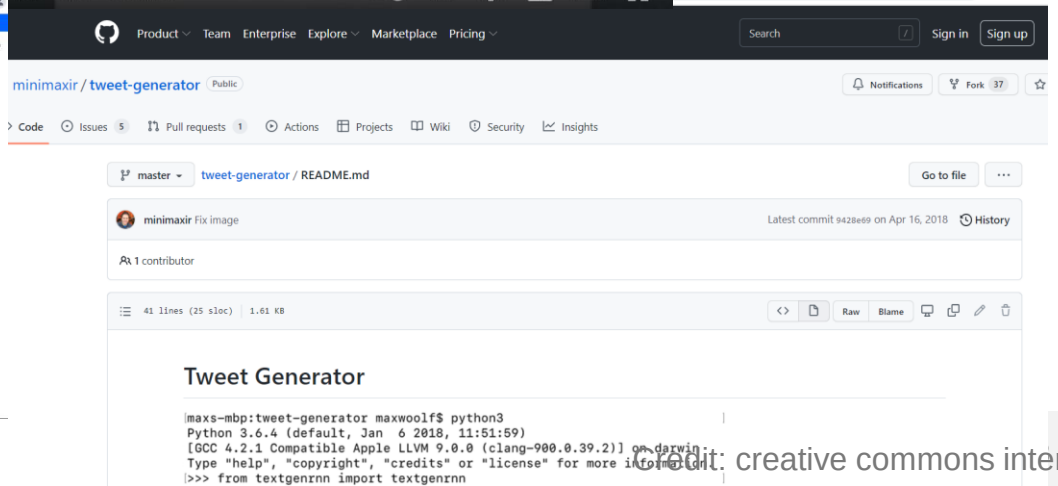


AWS Tutorial, TECHFRONT post









Creative Commons public domain

Credit: creative commons internet

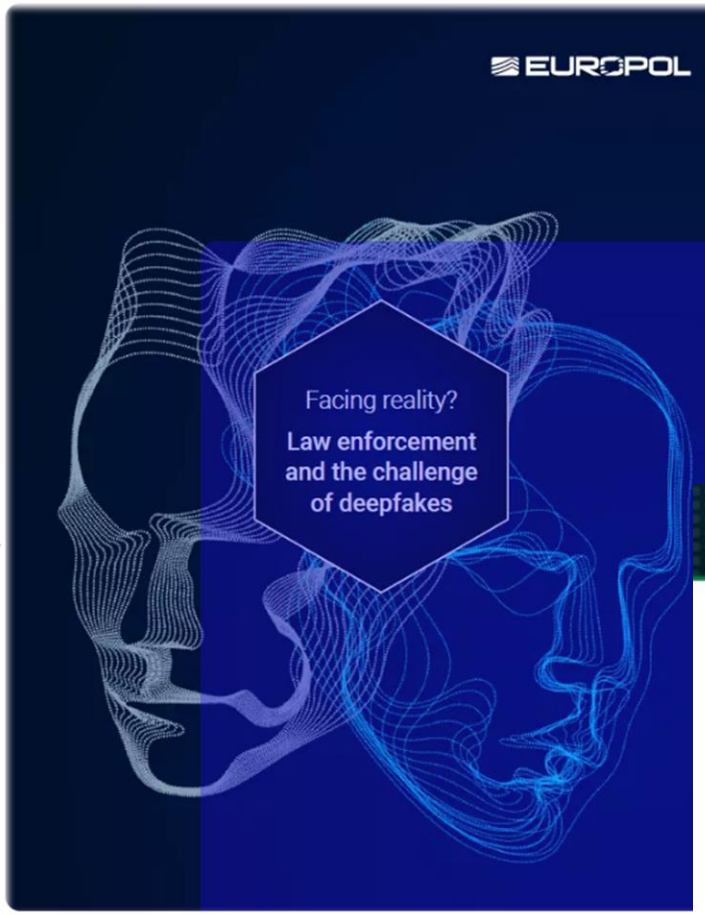
Deepfake warnings

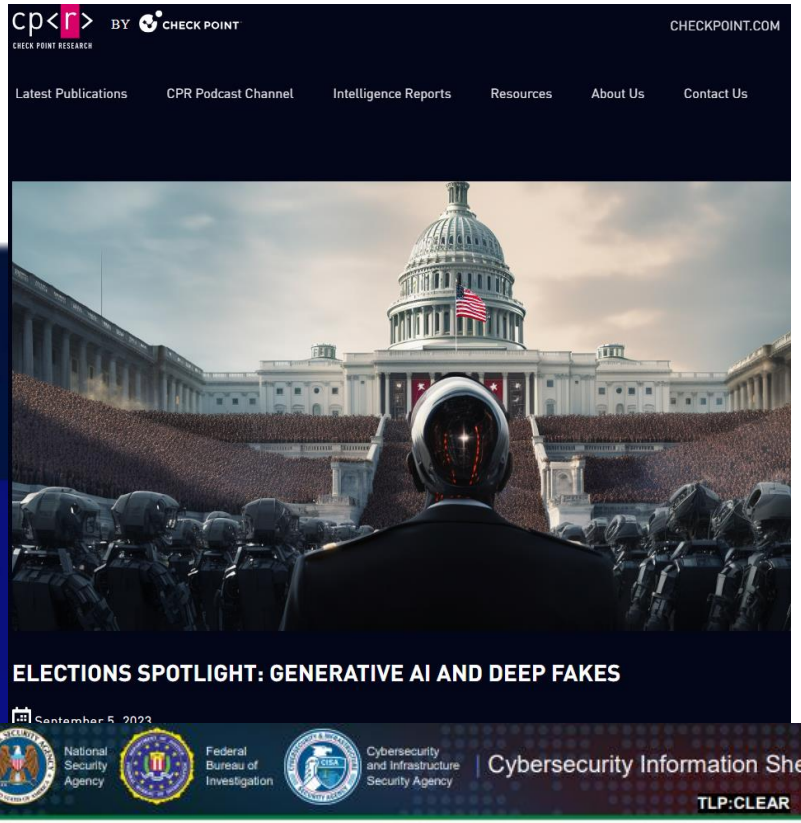
CEO impersonation for financial fraud
https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=1a968bc27559

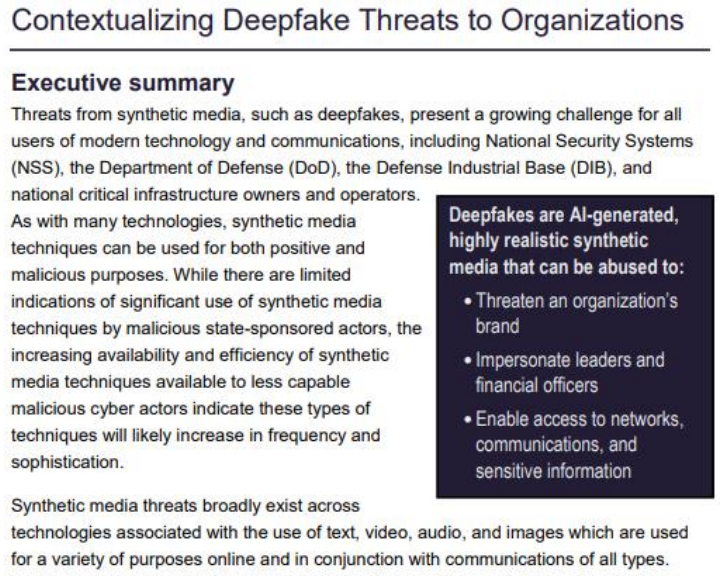
Europol warning use in organized crime
https://www.europol.europa.eu/media-press/newsroom/news/europol-report-finds-deepfake-technology-could-become-staple-tool-for-organised-crime

Check Point Research on GAI and Deepfakes
https://research.checkpoint.com/2023/elections-spotlight-generative-ai-and-deep-fakes/

CISA
https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF






Contextualizing Deepfake Threats to Organizations

Executive summary

Threats from synthetic media, such as deepfakes, present a growing challenge for all users of modern technology and communications, including National Security Systems (NSS), the Department of Defense (DoD), the Defense Industrial Base (DIB), and national critical infrastructure owners and operators. As with many technologies, synthetic media techniques can be used for both positive and malicious purposes. While there are limited indications of significant use of synthetic media techniques by malicious state-sponsored actors, the increasing availability and efficiency of synthetic media techniques available to less capable malicious cyber actors indicate these types of techniques will likely increase in frequency and sophistication.

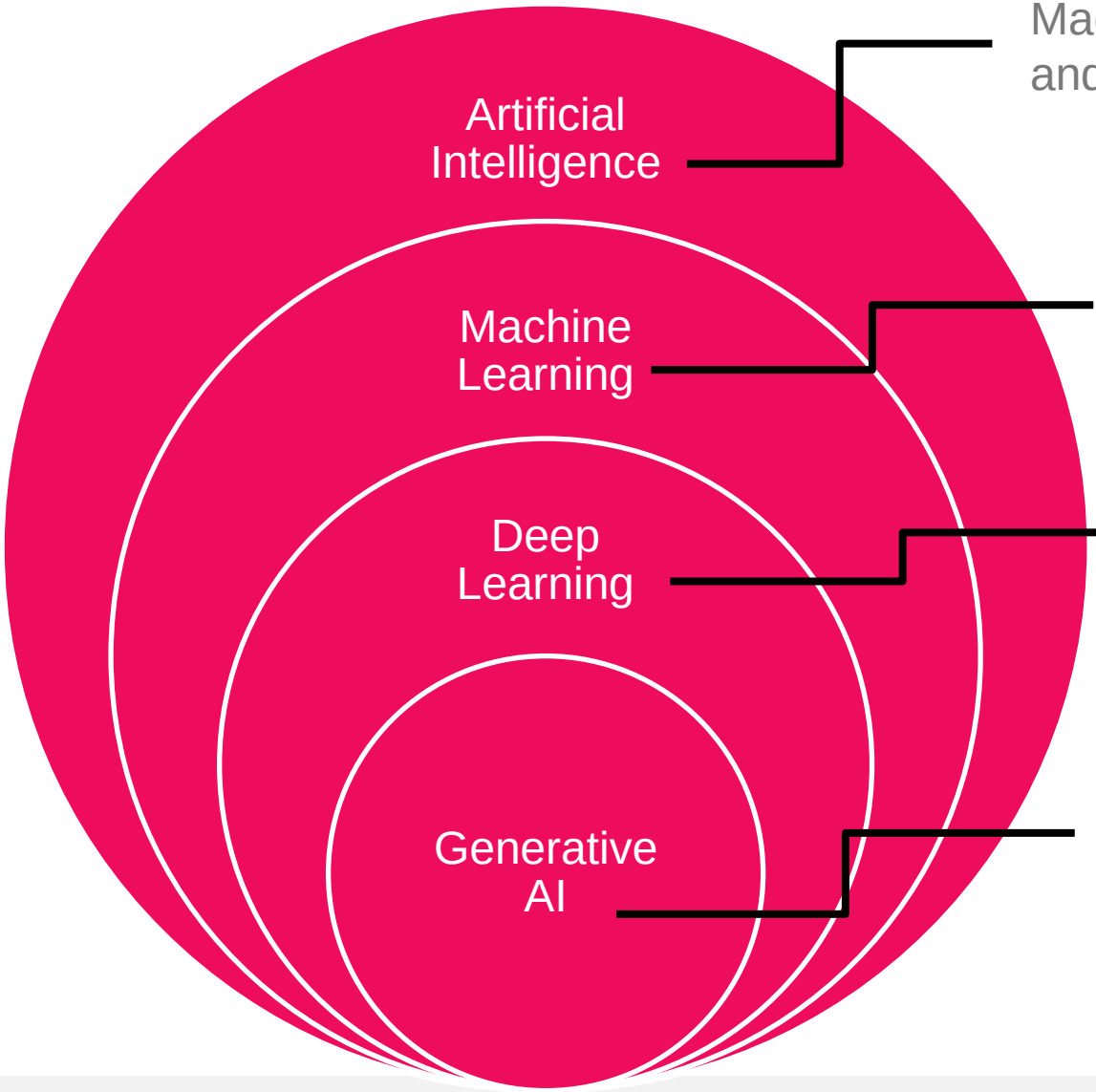
Deepfakes are AI-generated, highly realistic synthetic media that can be abused to:

- Threaten an organization's brand
- Impersonate leaders and financial officers
- Enable access to networks, communications, and sensitive information

Synthetic media threats broadly exist across technologies associated with the use of text, video, audio, and images which are used for a variety of purposes online and in conjunction with communications of all types.

Generative AI explosion with OpenAI and ChatGPT LLM 3.5 and LLM 4

Key terminologies in AI



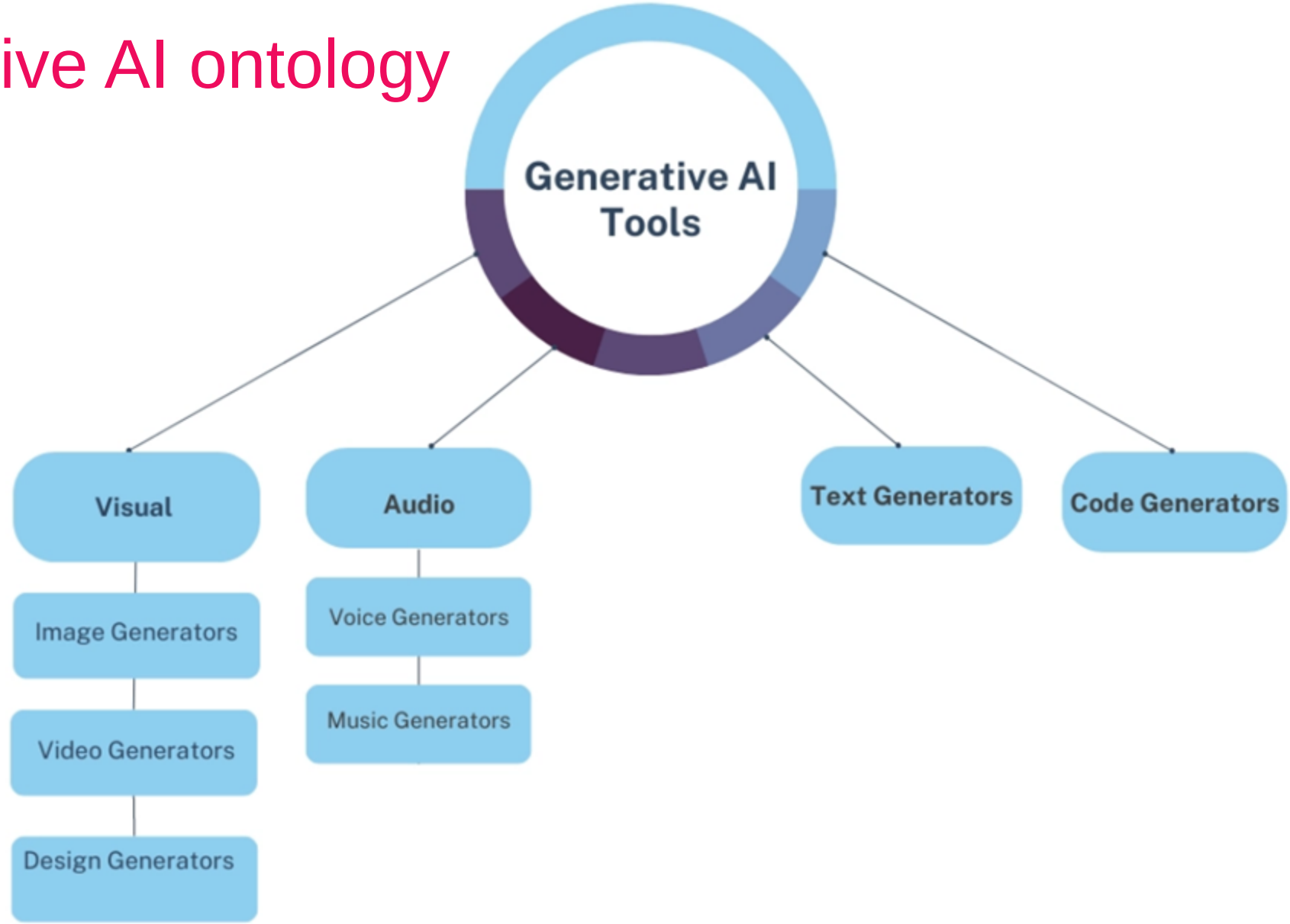
Machine that mimics human intelligence and behavior

Machine that learns from the data automatically w.r.t. a particular task and performance measure

Subset of ML, learning happens via algorithms inspired by neural networks in the human brain

Subset of DL, creating new content from existing ones using advanced algorithms

Generative AI ontology



Creative Commons public domain

Categories and platforms

Creative Commons public domain

Risk, rewards, tradeoffs

TIME SAVINGS: ChatGPT can generate responses quickly and efficiently, which can save time for both the user and the company using the technology.

CUSTOMIZATION: ChatGPT can be customized to fit the needs of the user, allowing for a personalized experience that can improve customer satisfaction and engagement.

SCALE: ChatGPT can handle a large volume of interactions at once, making it an effective solution for companies that need to handle high volumes of customer inquiries.

BIAS: ChatGPT is trained on a large amount of text data, which can contain biases, stereotypes, and offensive language. ChatGPT may perpetuate them in its responses and has potential to manipulate the user.

SAFETY: ChatGPT is not good for decision making. There is risk its responses will be inappropriate for crisis management or healthcare decisions.

ZERO EMPATHY: ChatGPT is an AI language model and lacks the ability to empathize with users. This may result in responses that are impersonal or unsympathetic, which could negatively impact user experience.

ETHICS BYPASS: ChatGPT API with third party software (Telegram) uses OpenAI's GPT-3 model and can circumvent ChatGPT ethics.

FRAGILE: ChatGPT requires extensive training to function effectively, which can be time-consuming and expensive for companies.

CONTEXT PROBLEMS AND BAD FOR DECISION MAKING: ChatGPT can generate responses based only on the information it has been given, it cannot understand the full context of a situation. This can lead to inaccurate or inappropriate responses.

Micki created with DALL•E and ChatGPT

Corporate, defender, adversary perspectives

The battle of AI

DEFENDER

Check Point has 40+ AI threat engines in prevention first architecture with ThreatCloud AI

Effective and efficient for Malware DNA genotyping, identification and blocking

Helps SOC analysts see attack vectors and landscape

Aid write and test good software code

Fix bugs in source code

Auto-write cybersecurity policies and controls based on GRC framework(s)

Gamify cybersecurity training

Mid-journey with DALL-E

ADVERSARY

Check Point Research saw major attacks created by ChatGPT since its inception

Create Deepfakes and bots

Malware writing for dummies (joke)

Phishing email creation for dummies (again a joke)

Easy to identify attack landscape based on vulnerabilities and find exploits to match (multiphased attacks)

Circumvent AI ethics of LLM GPT by using API with Telegram or other software integration

Adverary use of Gen AI

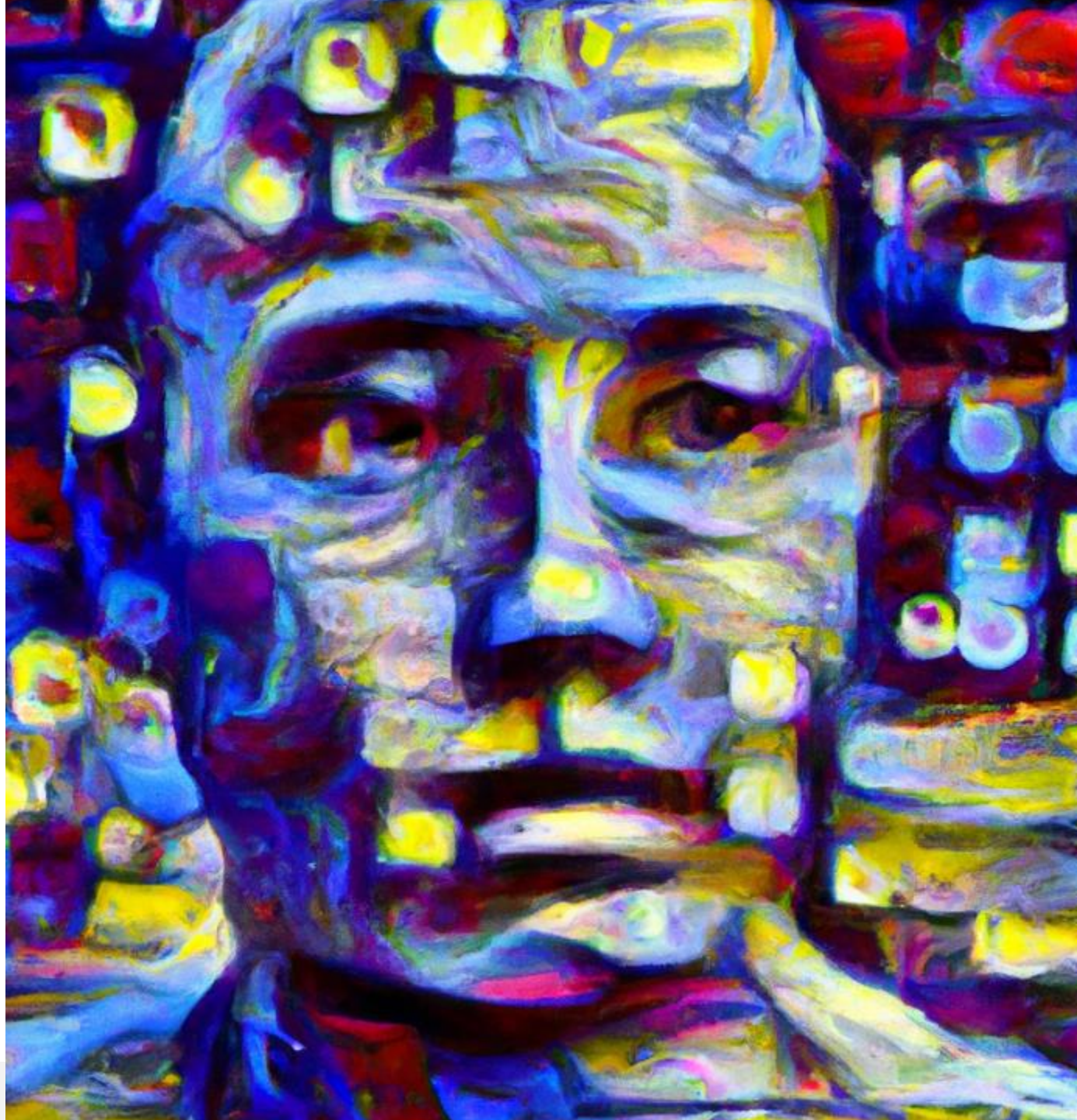
FraudGPT	Copycat hacking tool uses ChatGPT to create copycat tools for cyberattacks. Subscription based, seeks to use GAI for breakout attacks. $200 per month, up to $1,700 a year. Estimated 3000 users as of end of JUL 2023.
WormGPT *active favorite	**WormGPT leverages GPT LLM with specific focus on sophisticated phishing and spear phishing campaigns for business email compromise (BEC). Creates strategic and very convincing fake emails for use in large scale phishing and malware campaigns.**
PoisonGPT	For generating and spreading fake news and fake data regarding historical events. Used to manipulate, destabilize, and sway public opinion.
WolfGPT	Python-built alternative to ChatGPT. Superior evasion capabilities and the possibility to generate malicious content and advanced phishing attacks.
Evil GPT	Python based alternative to WormGPT.
DarkBART	Hackers are using to create sophisticated phishing campaigns, exploit vulnerabilities, create and spread malware.
XXXGPT **fake news?	**Offers hackers malicious subscription services from malware to botnets, RATs, infostealers, key loggers, and cryptostealers.**
DarkBERT	Criminal/cybercriminal GPT-based chatbot training on Dark Web. Based on Google Bard, uses Google Lens for images to create “criminal underground” knowledge base of cybercriminal and criminal TTPs.

Protect your data

DATA PRIVACY: API endpoint exposed by OpenAI should not retain or save any part of training data provided to it as part of the model fine-tuning/training process. No third party has access to the data shown to the model as a part of the training prompt by providing any kind of input to the exposed API endpoint.

DATA RETENTION AND LEAKAGE: has a default “data retention” period which requires the model to keep the training data for detect/prevent misuse of the API capabilities. Corporate custom data privacy agreements this retention period can be adjusted. Requires Mutual MOU then data will be scrubbed from the OpenAI systems. Protect data leakage by creating OpenAI data and model silos. OpenAI will simply silo off the requests/asks data independent of retention period, third parties will never have access or be able to extract your data by providing any input to the API.

Micki Boland created with DALL-E



https://chat.openai.com/chat

+ New chat

ChatGPT's Data Protecti

Generative AI Vendors Categor

Classical Mechanics Action Fun

Clear conversations

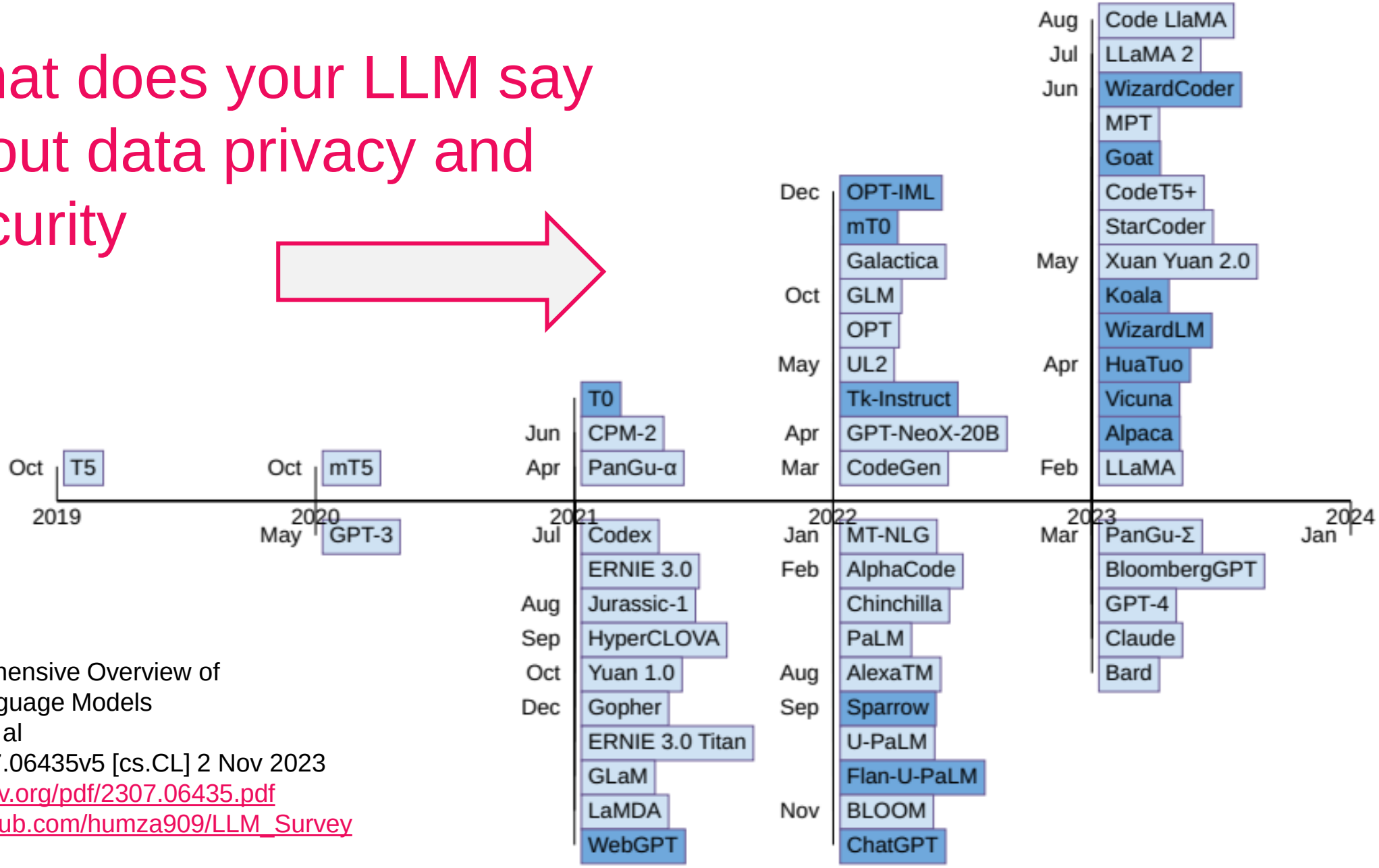
ChatGPT says?

As an AI language model, ChatGPT takes data protection and privacy very seriously, and we understand that these concerns are crucial for corporations. ChatGPT is committed to protecting corporate data and maintaining privacy in accordance with governance, risk, and compliance. Here are some of the ways we do this:

1. Compliance with Applicable Laws and Regulations: ChatGPT adheres to all applicable data protection and privacy laws and regulations, such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and other global privacy laws.
2. Robust Security Measures: We have implemented robust security measures to ensure the confidentiality, integrity, and availability of corporate data. This includes encryption of data at rest and in transit, network segmentation, firewalls, access controls, and regular security audits.
3. Access Control: ChatGPT limits access to corporate data to only authorized personnel who require access to perform their jobs. We also have strict password policies, multi-factor authentication, and monitoring of access to corporate data.
4. Data Protection by Design and Default: We implement data protection by design and default by incorporating privacy and data protection into the development and implementation of our AI models and services.
5. Data Minimization: ChatGPT minimizes the collection, storage, and processing of corporate data to only what is necessary for the provision of our services.
6. Transparency and Accountability: We are transparent about our data protection and privacy practices and provide our clients with regular reports on how their data is being used. We are also accountable for any data breaches and take appropriate action to mitigate any potential harm.

Overall, ChatGPT is committed to protecting corporate data and maintaining privacy in

What does your LLM say about data privacy and security



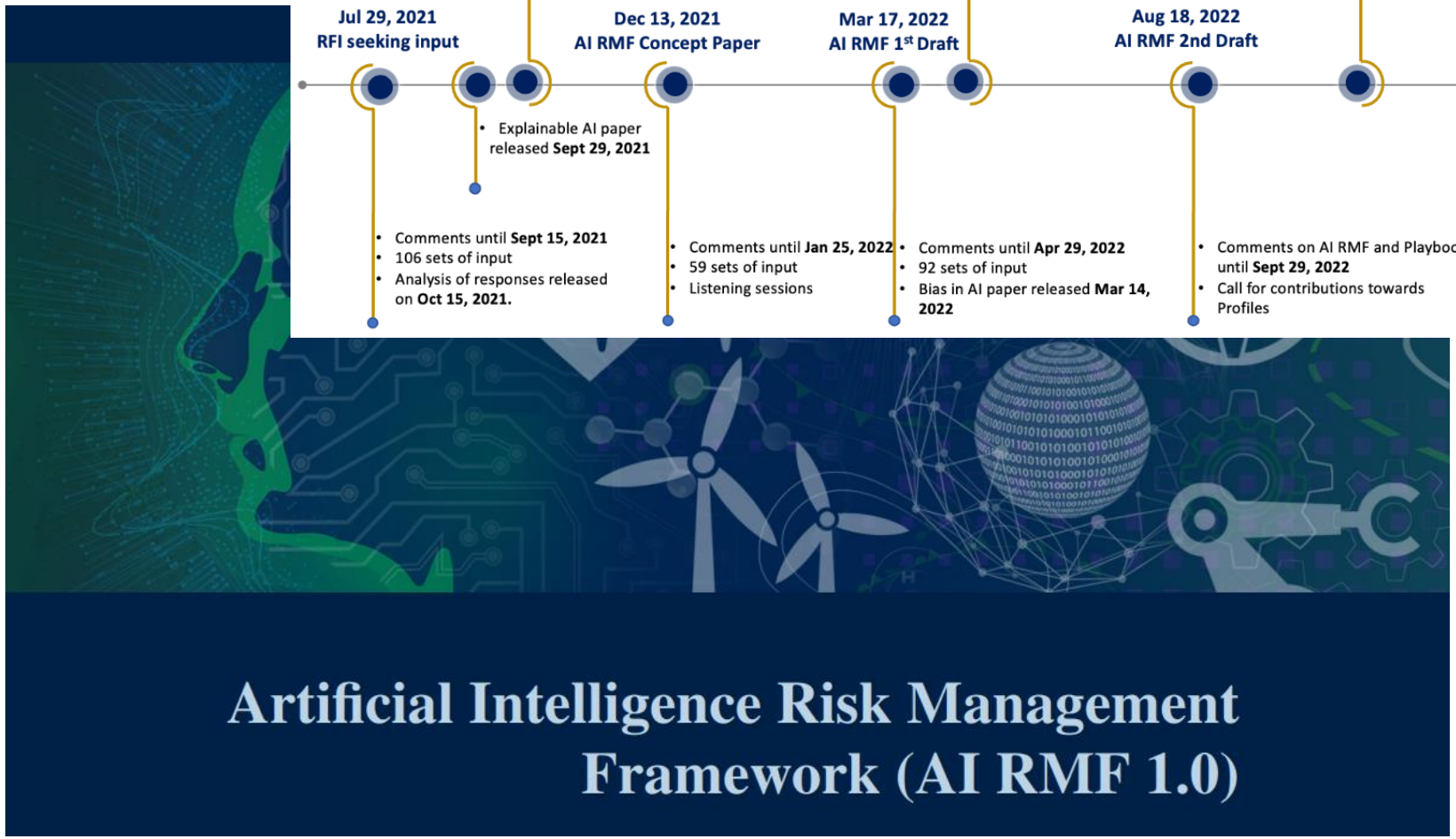
A Comprehensive Overview of Large Language Models
Humza, et al
arXiv:2307.06435v5 [cs.CL] 2 Nov 2023
https://arxiv.org/pdf/2307.06435.pdf
https://github.com/humza909/LLM_Survey

NIST AI RMF v1.0

NIST released on 26 JAN 2023 a 42 page Artificial Intelligence Risk Management Framework (AI RMF 1.0) and AI Risk Management Playbook

https://www.nist.gov/itl/ai-risk-management-framework

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

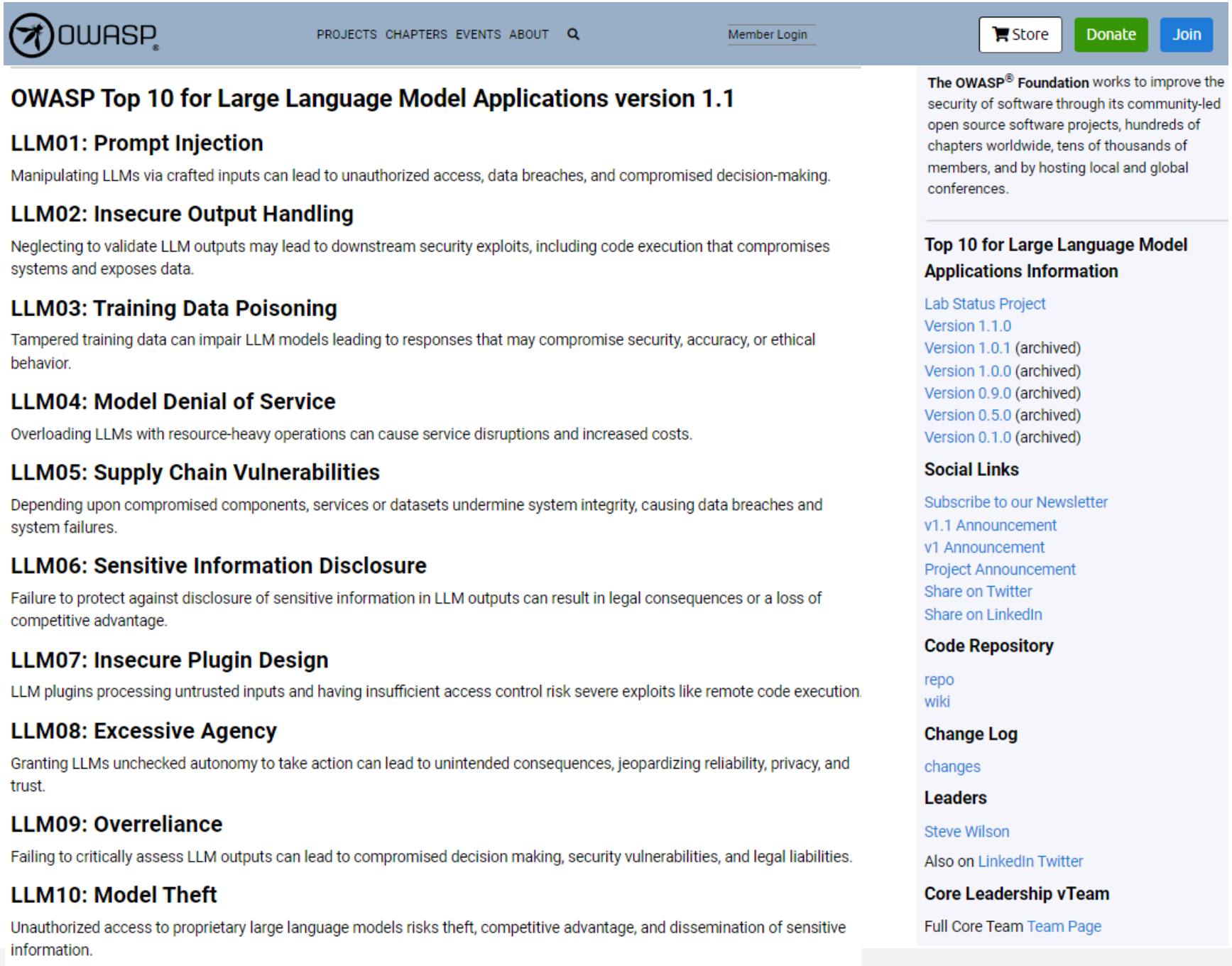


OWASP top 10 LLM risks

v1.1 is available

https://owasp.org/www-project-top-10-for-large-language-model-applications/

https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf



OWASP PROJECTS CHAPTERS EVENTS ABOUT Member Login Store Donate Join

OWASP Top 10 for Large Language Model Applications version 1.1

LLM01: Prompt Injection

Manipulating LLMs via crafted inputs can lead to unauthorized access, data breaches, and compromised decision-making.

LLM02: Insecure Output Handling

Neglecting to validate LLM outputs may lead to downstream security exploits, including code execution that compromises systems and exposes data.

LLM03: Training Data Poisoning

Tampered training data can impair LLM models leading to responses that may compromise security, accuracy, or ethical behavior.

LLM04: Model Denial of Service

Overloading LLMs with resource-heavy operations can cause service disruptions and increased costs.

LLM05: Supply Chain Vulnerabilities

Depending upon compromised components, services or datasets undermine system integrity, causing data breaches and system failures.

LLM06: Sensitive Information Disclosure

Failure to protect against disclosure of sensitive information in LLM outputs can result in legal consequences or a loss of competitive advantage.

LLM07: Insecure Plugin Design

LLM plugins processing untrusted inputs and having insufficient access control risk severe exploits like remote code execution.

LLM08: Excessive Agency

Granting LLMs unchecked autonomy to take action can lead to unintended consequences, jeopardizing reliability, privacy, and trust.

LLM09: Overreliance

Failing to critically assess LLM outputs can lead to compromised decision making, security vulnerabilities, and legal liabilities.

LLM10: Model Theft

Unauthorized access to proprietary large language models risks theft, competitive advantage, and dissemination of sensitive information.

The OWASP® Foundation works to improve the security of software through its community-led open source software projects, hundreds of chapters worldwide, tens of thousands of members, and by hosting local and global conferences.

Top 10 for Large Language Model Applications Information

Lab Status Project
Version 1.1.0
Version 1.0.1 (archived)
Version 1.0.0 (archived)
Version 0.9.0 (archived)
Version 0.5.0 (archived)
Version 0.1.0 (archived)

Social Links

Subscribe to our Newsletter
v1.1 Announcement
v1 Announcement
Project Announcement
Share on Twitter
Share on LinkedIn

Code Repository

repo
wiki

Change Log

changes

Leaders

Steve Wilson

Also on LinkedIn Twitter

Core Leadership vTeam

Full Core Team Team Page

OWASP Top 10 API risks
API security please!

SANS 2023 report only 50% organizations test their API!

Crucial! developers enriching API with LLM inventory and diagram out and API security test:
1. business use cases
2. logic flows

https://owasp.org/API-Security/editions/2023/en/0x11-t10/

Risk	Description
API1:2023 - Broken Object Level Authorization	APIs tend to expose endpoints that handle object identifiers, creating a wide attack surface of Object Level Access Control issues. Object level authorization checks should be considered in every function that accesses a data source using an ID from the user.
API2:2023 - Broken Authentication	Authentication mechanisms are often implemented incorrectly, allowing attackers to compromise authentication tokens or to exploit implementation flaws to assume other user's identities temporarily or permanently. Compromising a system's ability to identify the client/user, compromises API security overall.
API3:2023 - Broken Object Property Level Authorization	This category combines API3:2019 Excessive Data Exposure and API6:2019 - Mass Assignment, focusing on the root cause: the lack of or improper authorization validation at the object property level. This leads to information exposure or manipulation by unauthorized parties.
API4:2023 - Unrestricted Resource Consumption	Satisfying API requests requires resources such as network bandwidth, CPU, memory, and storage. Other resources such as emails/SMS/phone calls or biometrics validation are made available by service providers via API integrations, and paid for per request. Successful attacks can lead to Denial of Service or an increase of operational costs.
API5:2023 - Broken Function Level Authorization	Complex access control policies with different hierarchies, groups, and roles, and an unclear separation between administrative and regular functions, tend to lead to authorization flaws. By exploiting these issues, attackers can gain access to other users' resources and/or administrative functions.
API6:2023 - Unrestricted Access to Sensitive Business Flows	APIs vulnerable to this risk expose a business flow - such as buying a ticket, or posting a comment - without compensating for how the functionality could harm the business if used excessively in an automated manner. This doesn't necessarily come from implementation bugs.
API7:2023 - Server Side Request Forgery	Server-Side Request Forgery (SSRF) flaws can occur when an API is fetching a remote resource without validating the user-supplied URI. This enables an attacker to coerce the application to send a crafted request to an unexpected destination, even when protected by a firewall or a VPN.
API8:2023 - Security Misconfiguration	APIs and the systems supporting them typically contain complex configurations, meant to make the APIs more customizable. Software and DevOps engineers can miss these configurations, or don't follow security best practices when it comes to configuration, opening the door for different types of attacks.
API9:2023 - Improper Inventory Management	APIs tend to expose more endpoints than traditional web applications, making proper and updated documentation highly important. A proper inventory of hosts and deployed API versions also are important to mitigate issues such as deprecated API versions and exposed debug endpoints.
API10:2023 - Unsafe Consumption of APIs	Developers tend to trust data received from third-party APIs more than user input, and so tend to adopt weaker security standards. In order to compromise APIs, attackers go after integrated third-party services instead of trying to compromise the target API directly.

MITRE ATLAS

ATLAS™

The ATLAS Matrix below shows the general progression of attack tactics as column headers from left to right, with attack techniques organized below each tactic. & indicates a tactic or technique directly adapted from from ATT&CK. Click on the blue links to learn more about each item, or search and view more details about ATLAS tactics and techniques using the links in the top navigation bar.

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

TTPs adversarial use of AI, incorporated into MITRE ATT&CK Framework!
https://atlas.mitre.org/

THANK YOU!

Check Point cybersecurity use cases for Gen AI, LLM, API

Check Point cybersecurity for Gen AI, LLM and API

LLM and Generative AI Platforms Guardrails

- Quantum Application Control and URL Filtering

- URLF category for Artificial Intelligence: Application Signatures for Google Bard, ChatGPT, Microsoft Bing

- CloudGuard Network Security, CNAPP including SHIFTLEFT and CWP

Data Leak Protection over LLM

- Quantum and Harmony Connect DLP and Content Awareness to prevent leakage of sensitive data to GAI, add Harmony Mobile

API Security

- AppSec with schema validation

Impersonation fake .ai websites, phishing, watering holes and drivebys

- ThreatCloud AI, Threat Prevention AB, AV, IPS, DNS Reputation, DNS security, DGA

- Harmony Endpoint and Browser

Check Point cybersecurity for Gen AI, LLM and API

Check Point Quantum and Harmony Connect Data Loss prevention or Content Awareness to prevent leakage of sensitive data to Generative AI applications like chatGPT and Google Bard.

Create a security rule policy for Generative AI Applications and your sensitive data.

Video showing DLP for ChatGPT

https://community.checkpoint.com/t5/General-Topics/Preventing-leakage-of-sensitive-and-confidential-data-to/td-p/184234

Check Point cybersecurity for Gen AI, LLM and API

Test Tools

Building lab/test environments/tools: Kali, Burp Suite, FoxyProxy, Postman, OWASP Zap, Google hacking, Shodan, Wfuzz

Passive recon for finding exposed endpoints, fuzzing, detecting anomalies

LLM hacking, evasion, data exposure testing and demonstrate protections

API hacking schema (RESTful), misconfigurations, rate limiting, data exfil SQL and NoSQL injection, cross-API (XAS), cross-site scripting (XSS), and demonstrate protections

Check Point cybersecurity for Gen AI, LLM and API

AWS VPC and EC2 instances, containers, functions

AWS Bedrock

Check Point Security Management

CloudGuard NGTP Gateway ingress / egress protection

Test AWS Linux box with these tools: Kali, Burp Suite, FoxyProxy, Postman, OWASP Zap, Shodan, Wfuzz

Kong API Gateway

Openappsec agent deployed on Kong Gateway for API discovery, bot protection, schema validation, rate limiting

OpenAI API key for API

OpenAI ChatGPT 3.5 and test 4

CloudBot (for swarm SNS notifications)

GitHub new repo

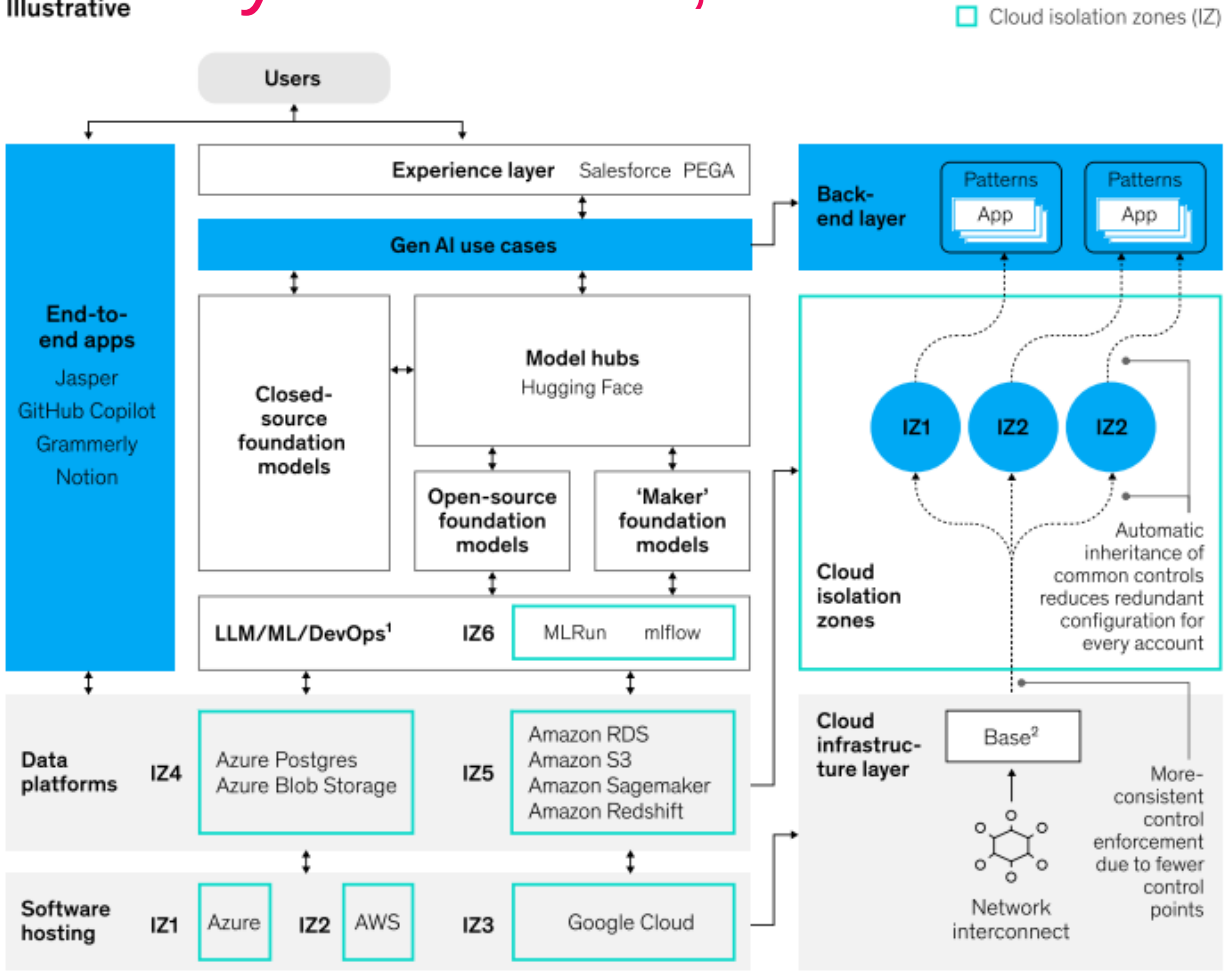
Harmony Mobile for test devices using Slack

Check Point cybersecurity for Gen AI, LLM and API

Exhibit 25

The right cloud foundation for generative AI involves an architecture connecting the back end, data, and cloud infrastructure

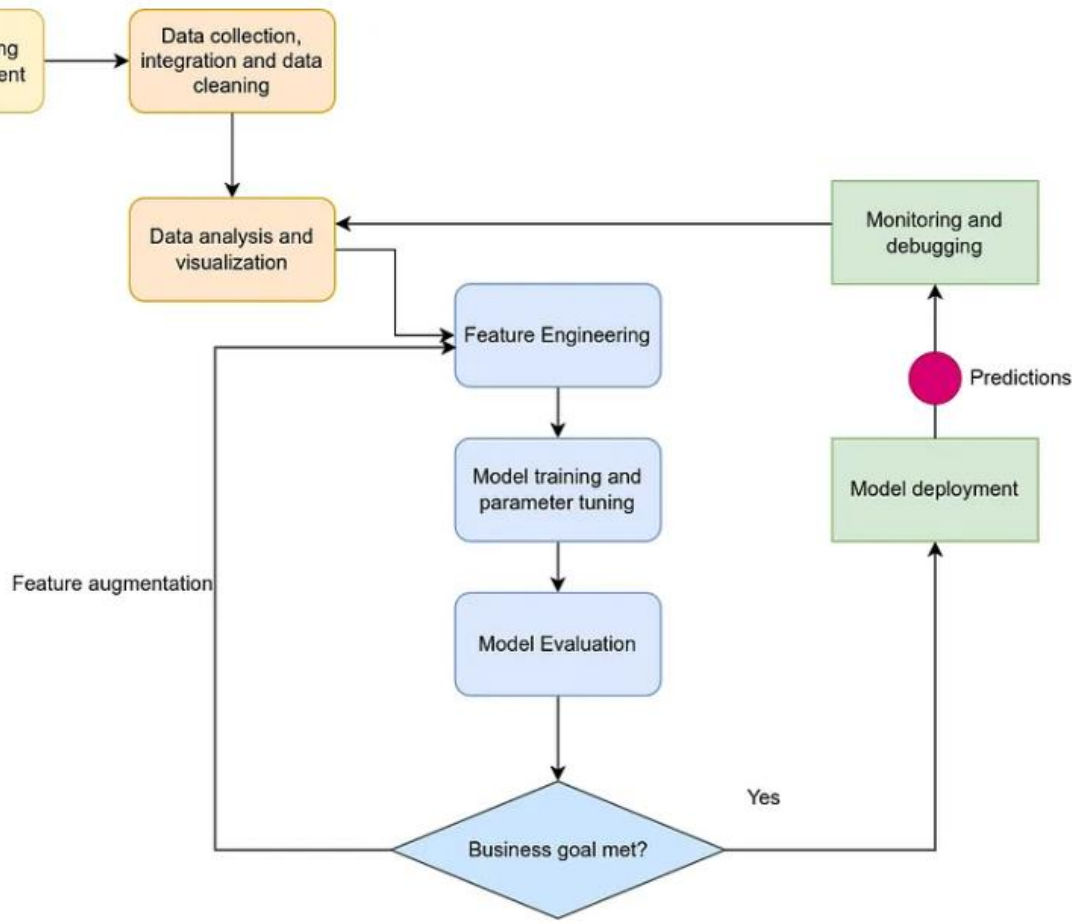
Illustrative



[1]LLM = large language model; ML = machine learning.
[2]A single primary base services most isolation zones, but additional bases (eg, an integration base for M&A activities) can also exist in the architecture.

1. **Understand business resiliency requirements for key journeys.** Technology and business leaders should jointly identify the most-critical business journeys and their associated applications. Applications can be categorized into four levels: mission critical, business critical, business operations, and administrative (Exhibit 26).

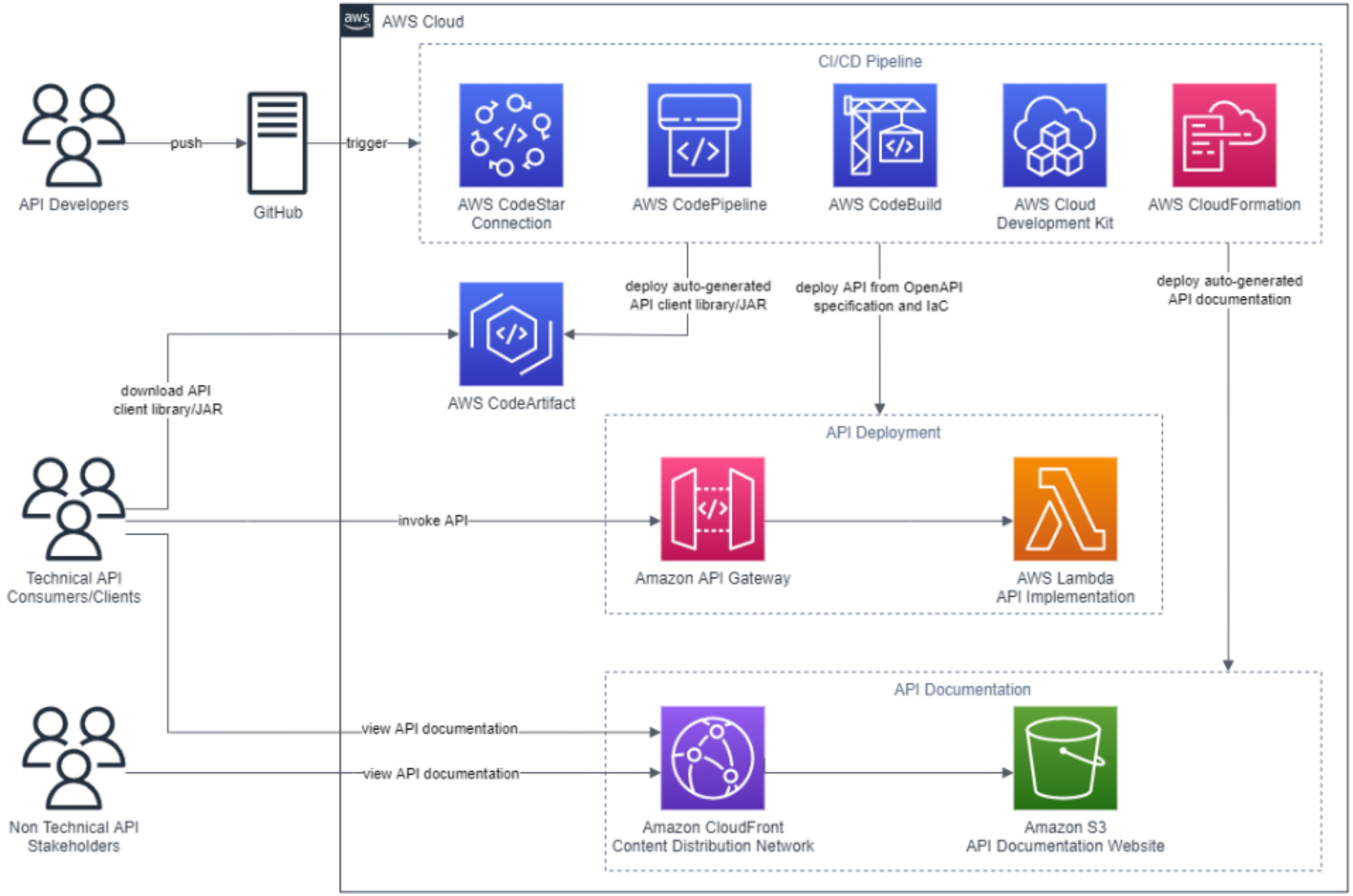
Real world ML process



Machine learning process — Image by Author

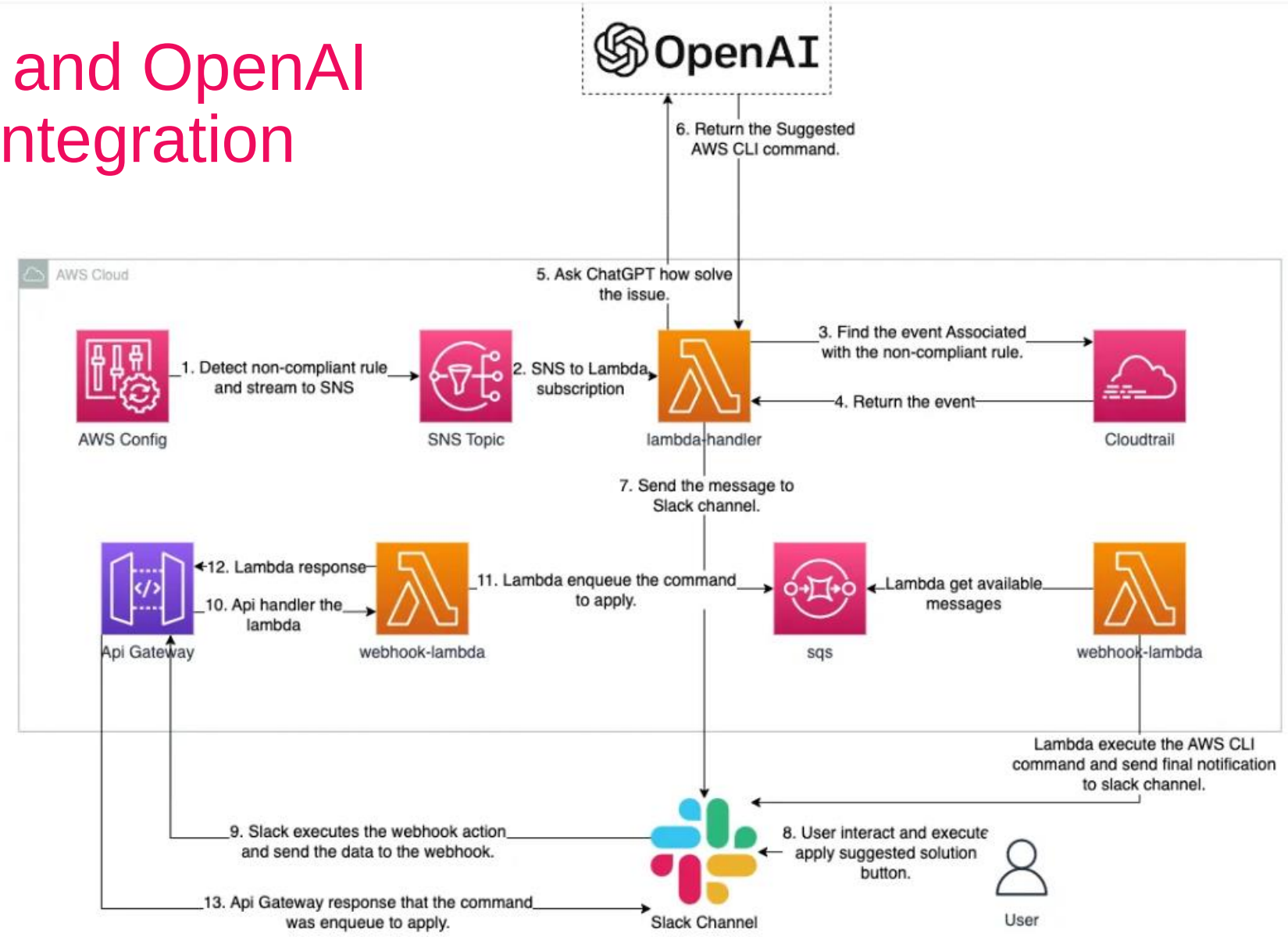
https://towardsdatascience.com/how-to-design-an-mlops-architecture-in-aws-67ee9843a430

Use case AWS and OpenAPI (Swaggerhub) integration



https://aws.amazon.com/blogs/devops/deploy-and-manage-openapi-swagger-restful-apis-with-the-aws-cloud-development-kit/

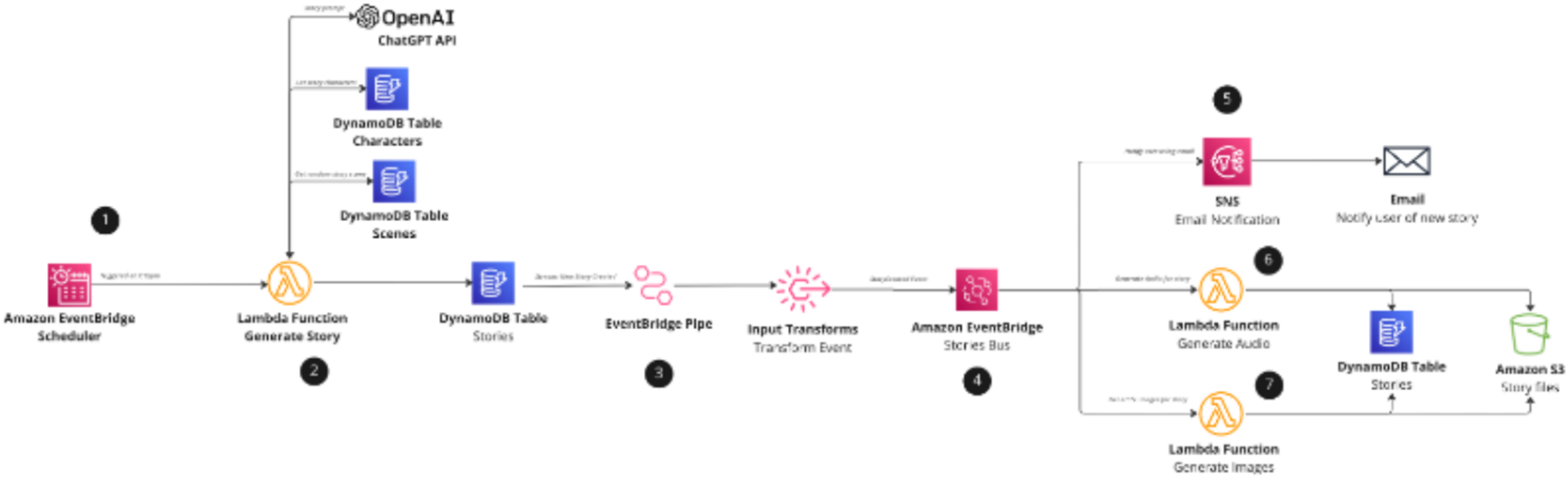
Use case AWS and OpenAI Slack Chatbot integration



https://medium.com/globant/integrating-chatgpt-api-with-aws-config-eb7633723b3

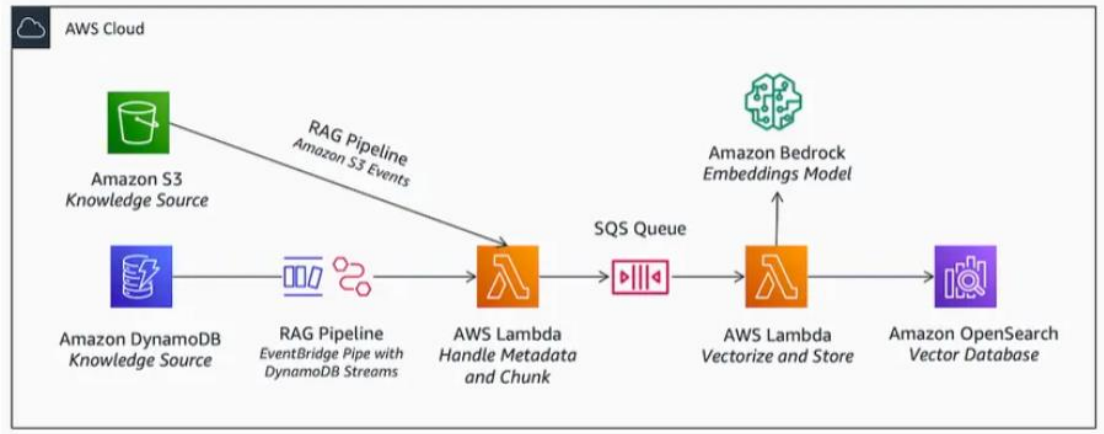
https://github.com/4l3j4ndr0/aws-config-chatGPT

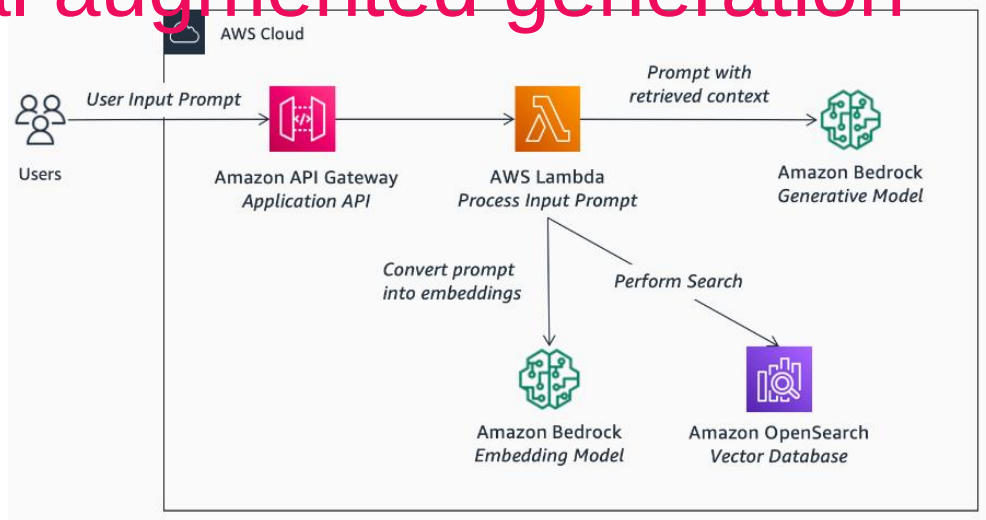
Use case AWS and ChatGPT API integration sgtory application with Lambda schedule

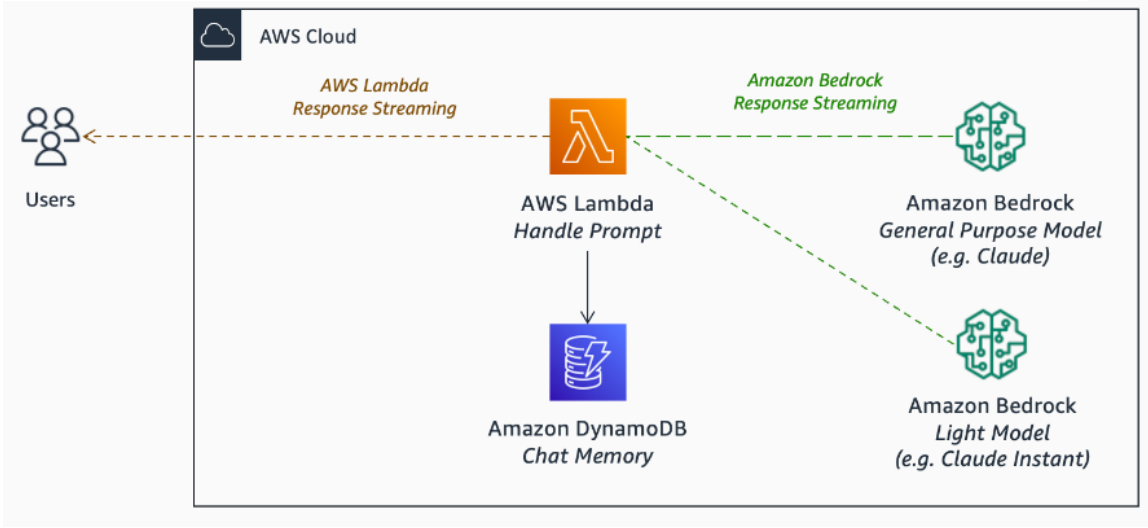


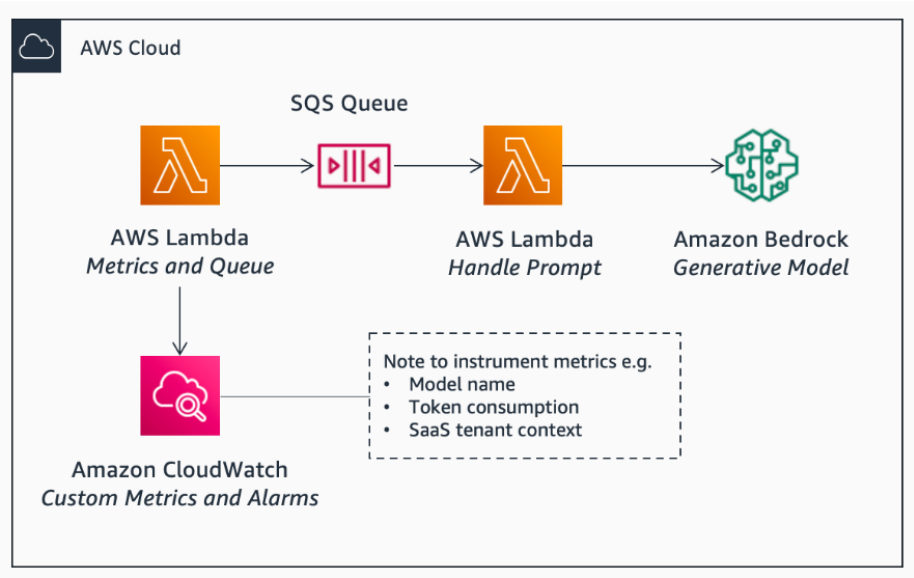
https://aws.amazon.com/blogs/compute/implementing-an-event-driven-serverless-story-generation-application-with-chatgpt-and-dall-e/

Use case AWS Bedrock AI retrieval augmented generation (RAG)









https://community.aws/posts/build-generative-ai-applications-with-amazon-bedrock

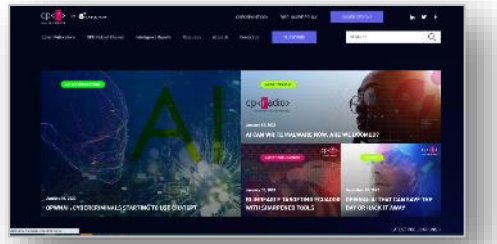
Check Point Research and CISO podcast resources

Check Point Research (CPr)

- cp<r> blog for technical and research publications.
- Live Threat Map - online
- Weekly threat intelligence reports

  cp<r> or subscribe
- Biannual “Cyber Trend Report”

  cp<r> or publications (look for the interactive)



Check Point Research (CPr)

- CP corporate blog for “monthly top malware”and other reports and publications.
- cp<radio> podcasts
- Customized intelligence reports
- cp<r> twitter - _CPResearch_
- CISO Secrets podcast