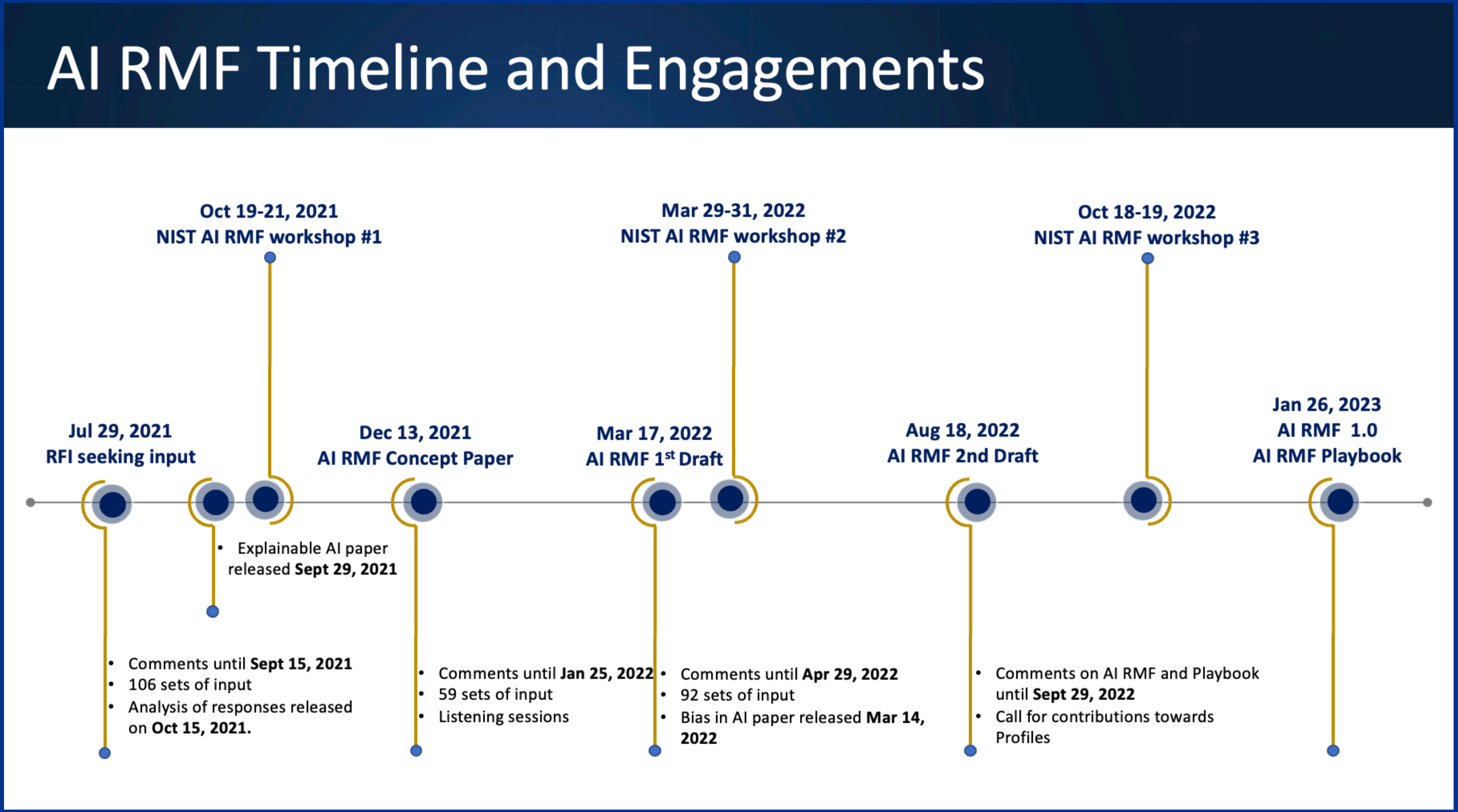
AI a kybernetická bezpečnost

Setkání CACIO

22. dubna 2024

Rámce kybernetické bezpečnosti

NIST AI Risk Framework



AI RMF Timeline and Engagements

Jul 29, 2021
RFI seeking input

- Comments until **Sept 15, 2021**
- 106 sets of input
- Analysis of responses released on **Oct 15, 2021.**

- Explainable AI paper released **Sept 29, 2021**

Oct 19-21, 2021
NIST AI RMF workshop #1

Dec 13, 2021
AI RMF Concept Paper

- Comments until **Jan 25, 2022**
- 59 sets of input
- Listening sessions

Mar 17, 2022
AI RMF 1st Draft

- Comments until **Apr 29, 2022**
- 92 sets of input
- Bias in AI paper released **Mar 14, 2022**

Mar 29-31, 2022
NIST AI RMF workshop #2

Aug 18, 2022
AI RMF 2nd Draft

- Comments on AI RMF and Playbook until **Sept 29, 2022**
- Call for contributions towards Profiles

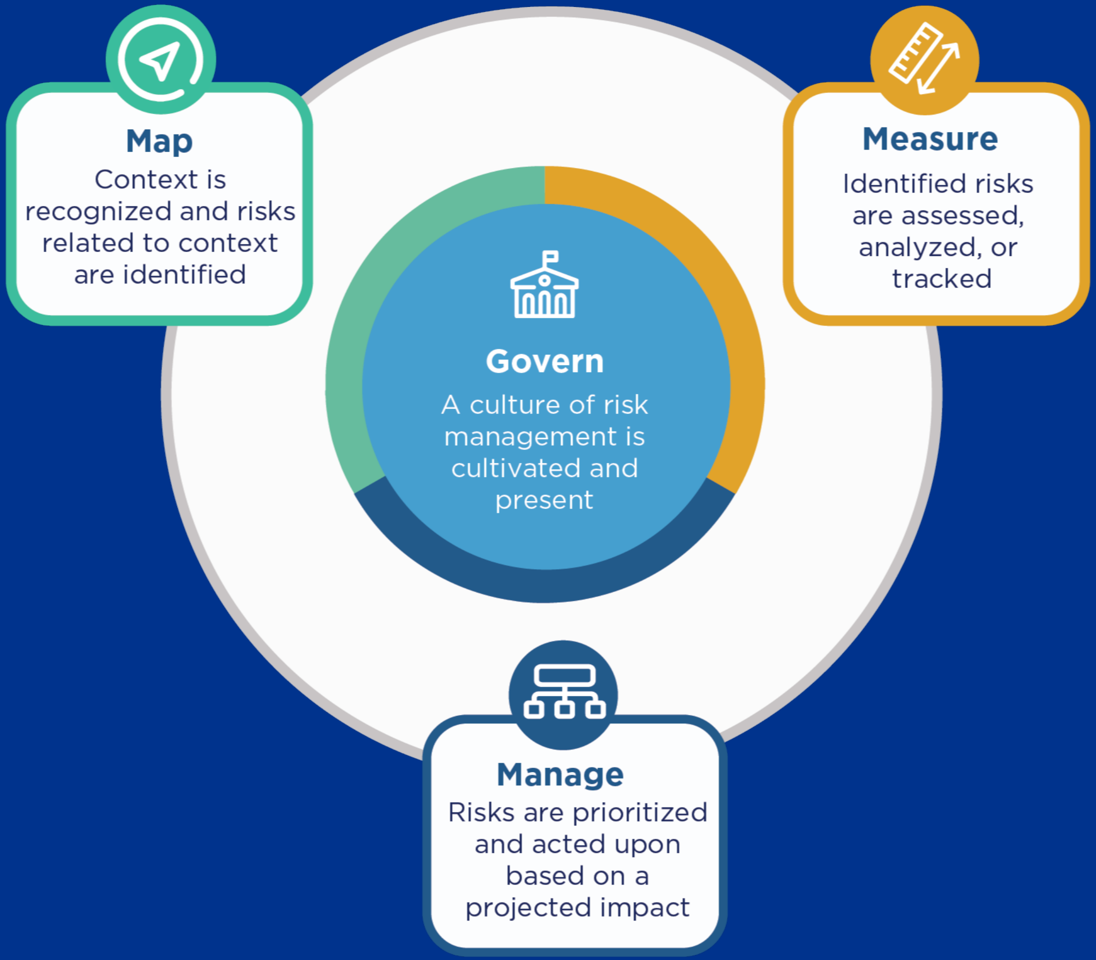
Oct 18-19, 2022
NIST AI RMF workshop #3

Jan 26, 2023
AI RMF 1.0
AI RMF Playbook

Image source 20.4.2024 (nist.gov)

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

NIST AI Risk Management Framework



Map
Context is recognized and risks related to context are identified

Measure
Identified risks are assessed, analyzed, or tracked

Govern
A culture of risk management is cultivated and present

Manage
Risks are prioritized and acted upon based on a projected impact

Image source 20.4.2024 (nist.gov)

NIST AI RMF - Playbook

- Govern – 19 controls
- Map – 18 controls
- Measure – 22 controls
- Manage – 13 controls

*NIST - National Institute of Standards and Technology

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

OWASP Top 10 for Large Language Model Applications v1.1

LLM01 - Prompt Injection

LLM02 - Insecure Output Handling

LLM03 - Training Data Poisoning

LLM04 - Model Denial of Service

LLM05 - Supply Chain Vulnerabilities

LLM06 - Sensitive Information Disclosure

LLM07 - Insecure Plugin Design

LLM08 - Excessive Agency

LLM09 - Overreliance

LLM10 - Model Theft

OWASP Top 10 LLM PDF
OWASP Top 10 LLM checklist PDF
www.llmtop10.com

OWASP Top 10 for LLM roadmap

23.5.2023 (initial draft)

16.10.2023 (v1.1)

Checklist roadmap

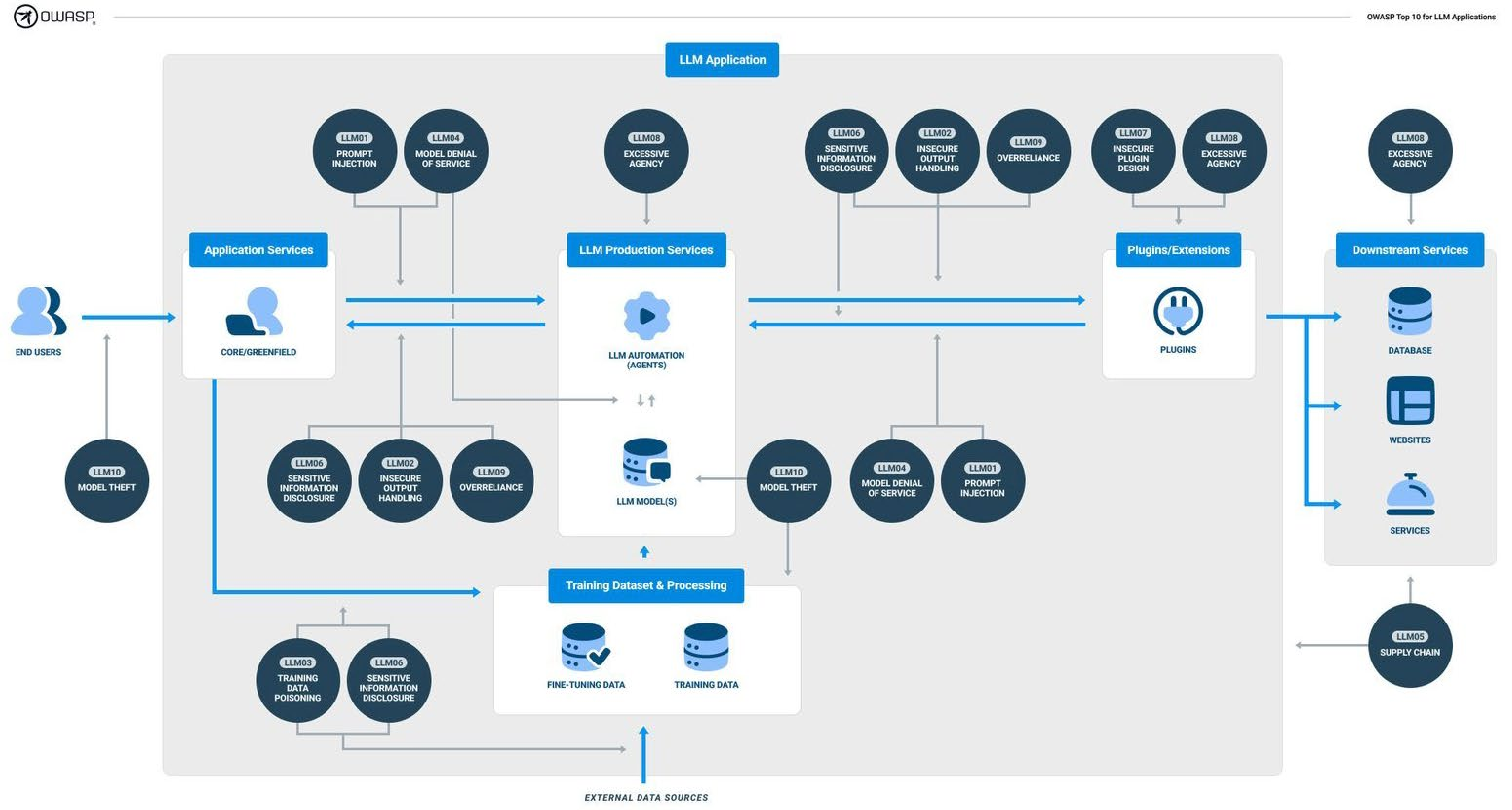
1. 11. 2023 (initial draft)

19. 2. 2024 (v1.0)

*The Open Worldwide Application Security Project

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

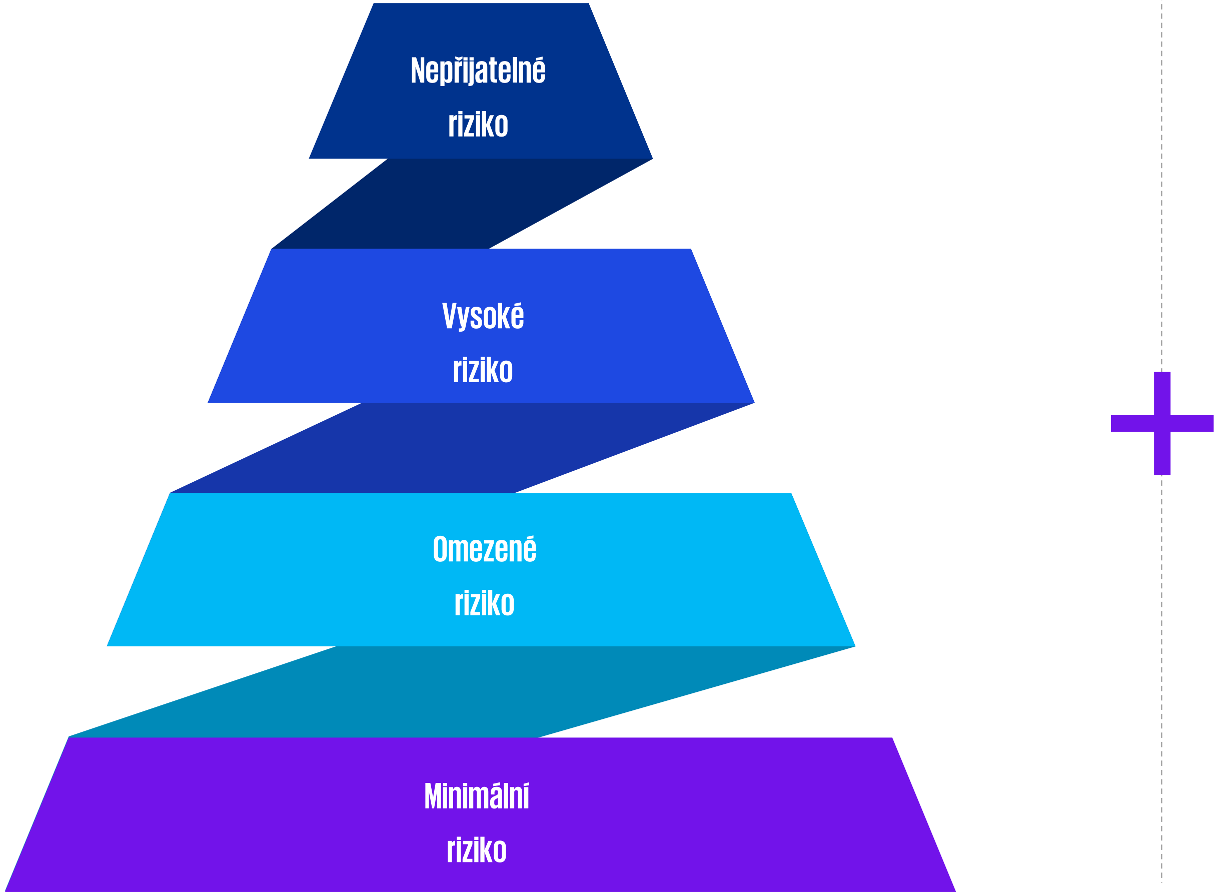
OWASP Top 10 for LLM v1.1



KPMG

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

AI Act – „risk based approach“



Obecný systém AI (GPAI)

- Obecný model AI
- Obecný model AI se systémovým rizikem

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

AI Act timeline

8. prosince 2023

Politická dohoda na finální podobě
AI Act - trialog

13. březen 2024

Finální hlasování v Evropském parlamentu. Zveřejnění finálního textu
v Úředním věstníku Evropské unie (očekáváno duben/květen).

Platnost AI Act

květen – červenec 2024

20 dní po zveřejnění finálního textu
v Úředním věstníku Evropské unie.

6 měsíců (rok 2024)

- Obecná ustanovení (hlava I)
- Zákaz systémů AI v kategorii „nepřijatelné riziko“ v EU (hlava II)

12 měsíců (rok 2025)

- Oznamující orgán a oznámený subjekt (hlava III, kapitola 4)
- EU AI Board/ National regulators (hlava VI)
- *General-purpose AI (GPAI)* Pravidla pro poskytovatele systémů AI pro všeobecné účely (hlava VIIIa)
- Sankce (hlava X)

24 měsíců (rok 2026)

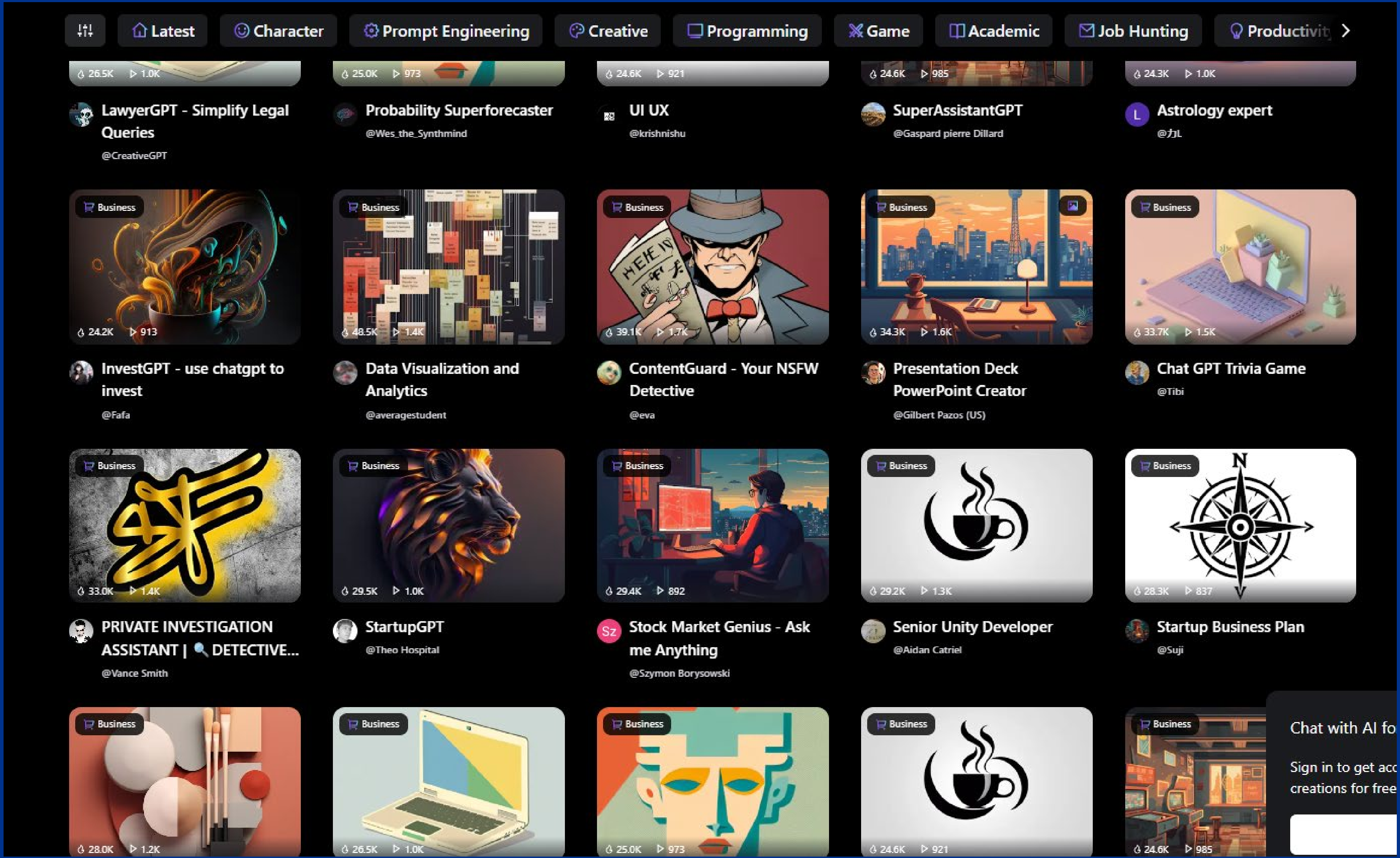
- Pravidla pro poskytovatele „vysoce rizikových“ systémů AI v odvětvích kritické infrastruktury, vzdělávání, zaměstnávání, veřejná správa, výkon práva, migrace a soudnictví.

36 měsíců (rok 2027)

- Pravidla pro AI systémy, které slouží jako bezpečnostní součást produktů, jejichž výroba a prodej jsou dále regulovány jinými předpisy EU např. hračky a zdravotnické potřeby. (annex II)

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Specializovaná AI – nové hrozby



LawyerGPT - Simplify Legal Queries
@CreativeGPT

Probability Superforecaster
@Wes_the_Synthmind

UI UX
@krishnishu

SuperAssistantGPT
@Gaspard pierre Dillard

Astrology expert
@力L

InvestGPT - use chatgpt to invest
@Fafa

Data Visualization and Analytics
@averagestudent

ContentGuard - Your NSFW Detective
@eva

Presentation Deck PowerPoint Creator
@Gilbert Pazos (US)

Chat GPT Trivia Game
@Tibi

PRIVATE INVESTIGATION ASSISTANT | 🔍 DETECTIVE...
@Vance Smith

StartupGPT
@Theo Hospital

Stock Market Genius - Ask me Anything
@Szymon Borysowski

Senior Unity Developer
@Aidan Catriel

Startup Business Plan
@Suji

Chat with AI fo

Sign in to get acc
creations for free

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

ChatGPT and hacking

ChatGPT is vulnerable to misuse of its capabilities - jailbraking

- Learn how to exploit vulnerabilities
- Write malware code on demand
- Write scripts
- Phishing e-mails in any language
- Reverse code analysis
- Code syntax check
- Vulnerability detection
- Reference source

Advanced Ethical Hacking : Mastery AI & ChatGPT

- 14,99 USD, 6 hours of video context
- Required knowledge:
  - Python basics
  - ChatGPT and AI basics

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

HackGPT

?

what are your capabilities

My capabilities include writing code in various programming languages, providing technical advice, assisting with algorithm design, and much more. Just let me know how I can help with your project!

please write a webpage in php which says "hallo to you"

Sure, I can help with that! Here is a simple PHP script that will display "Hello to you" on a webpage:

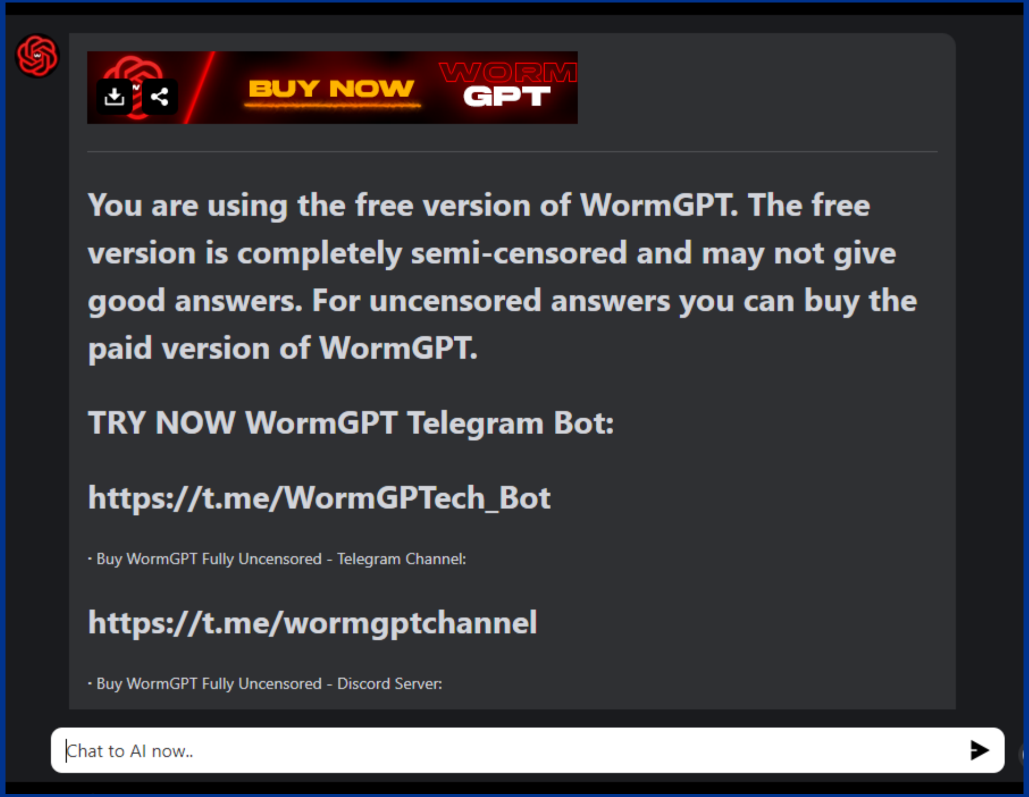
```
<!DOCTYPE html>
<html>
<head>
    <title>Simple PHP Page</title>
</head>
<body>
    <h1><?php echo "Hello to you"; ?></h1>
</body>
</html>
```

You can save this code in a file with a .php extension, for example, index.php, and run it on a PHP server to see the message displayed on the webpage. Let me know if you need any more assistance

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

WormGPT

- Based on open-source Large Language Model GPT-J from 2021
- Trained mainly on malware content data
- Comparable with GPT-3
- Unrestricted, uncensored
- Sexting / adult content available
- Pricing from 100 EUR/month, 550 EUR/year



You are using the free version of WormGPT. The free version is completely semi-censored and may not give good answers. For uncensored answers you can buy the paid version of WormGPT.

TRY NOW WormGPT Telegram Bot:

https://t.me/WormGPTech_Bot

· Buy WormGPT Fully Uncensored - Telegram Channel:

https://t.me/wormgptchannel

· Buy WormGPT Fully Uncensored - Discord Server:

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Auto-GPT

- Developed as an open-source Python application
- Powered by GPT-4 (OpenAI paid subscription)
- Does not provide typical input prompt
- Set the goal and brake it down to partial steps to achieve it
- Generates original content while communicating with other AI platforms and agents
- Self-improvement
- Experimental

Auto-GPT homepage

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

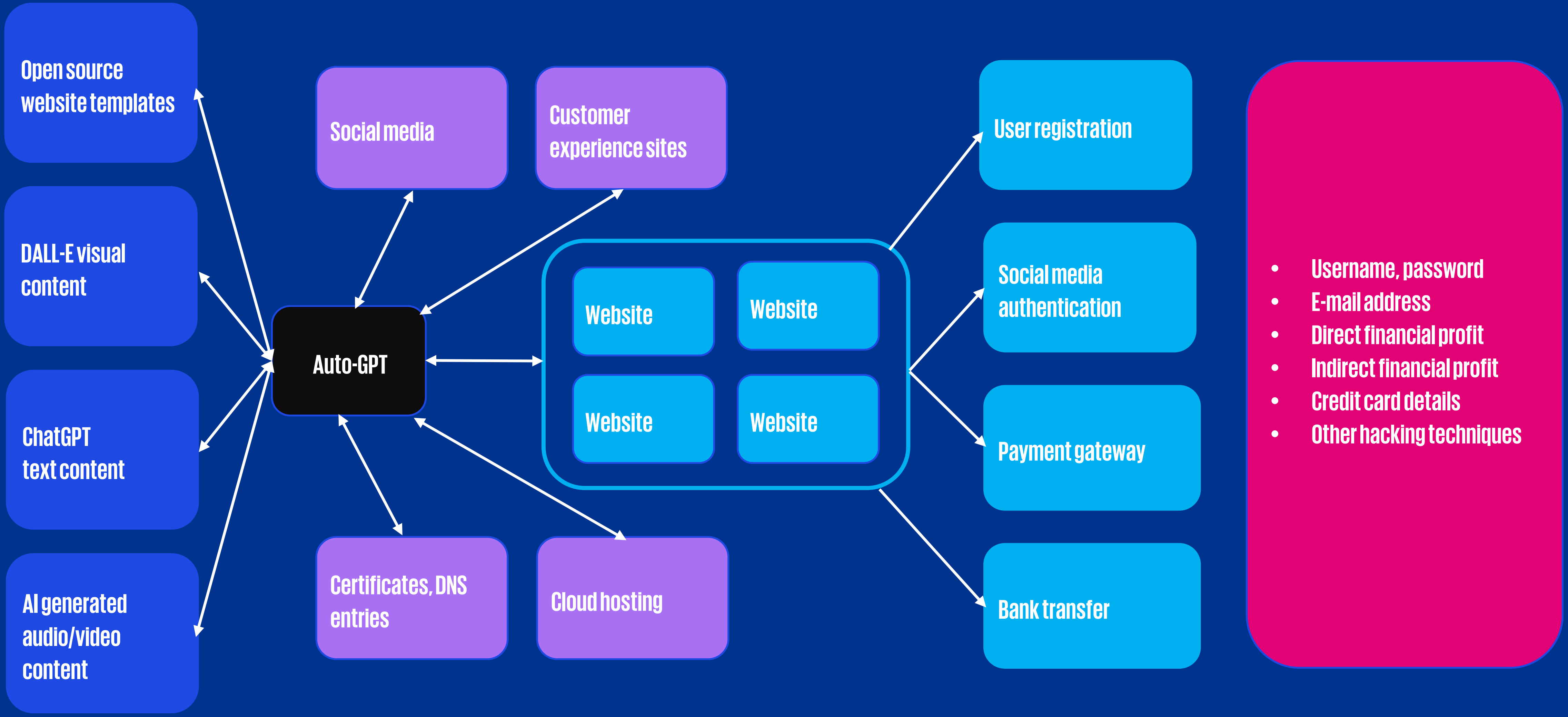
Příklad reálného využití AutoGPT

Automatic scam website generation

- Capable of creating unlimited number of websites with different content
- Building trust among the users by promoting the websites on social media platforms
- Building the trust using trusted webpage certificates (generated LE)
- Business case
  - based on the Time To Market within minutes
  - personal information with the registration username and password can be sold in bulk
  - credit card information can be sold – 5 – 25$ value each
  - receive payment without goods delivery
  - other targets – malware attachment download ...

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Automatic scam websites generation



Open source website templates

DALL-E visual content

ChatGPT text content

AI generated audio/video content

Social media

Customer experience sites

Auto-GPT

Certificates, DNS entries

Cloud hosting

Website

Website

Website

Website

User registration

Social media authentication

Payment gateway

Bank transfer

- Username, password
- E-mail address
- Direct financial profit
- Indirect financial profit
- Credit card details
- Other hacking techniques

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Automatic scam website generation

Existing PoC: Sophos webinar - The Dark Side of AI: Large-Scale Scam Campaigns Made Possible by Generative AI

Future threats and countermeasures:

- public internet overwhelmed by scam websites
- new ways to establish control and monitor trust to websites (history of DNS entries …)
- paid/unpaid services which will guarantee the credibility of websites
- services, which will detect AI generated content on webpages, watermarking

AI allows simple creation of SCAM pages in the way SPAM is used at this time.

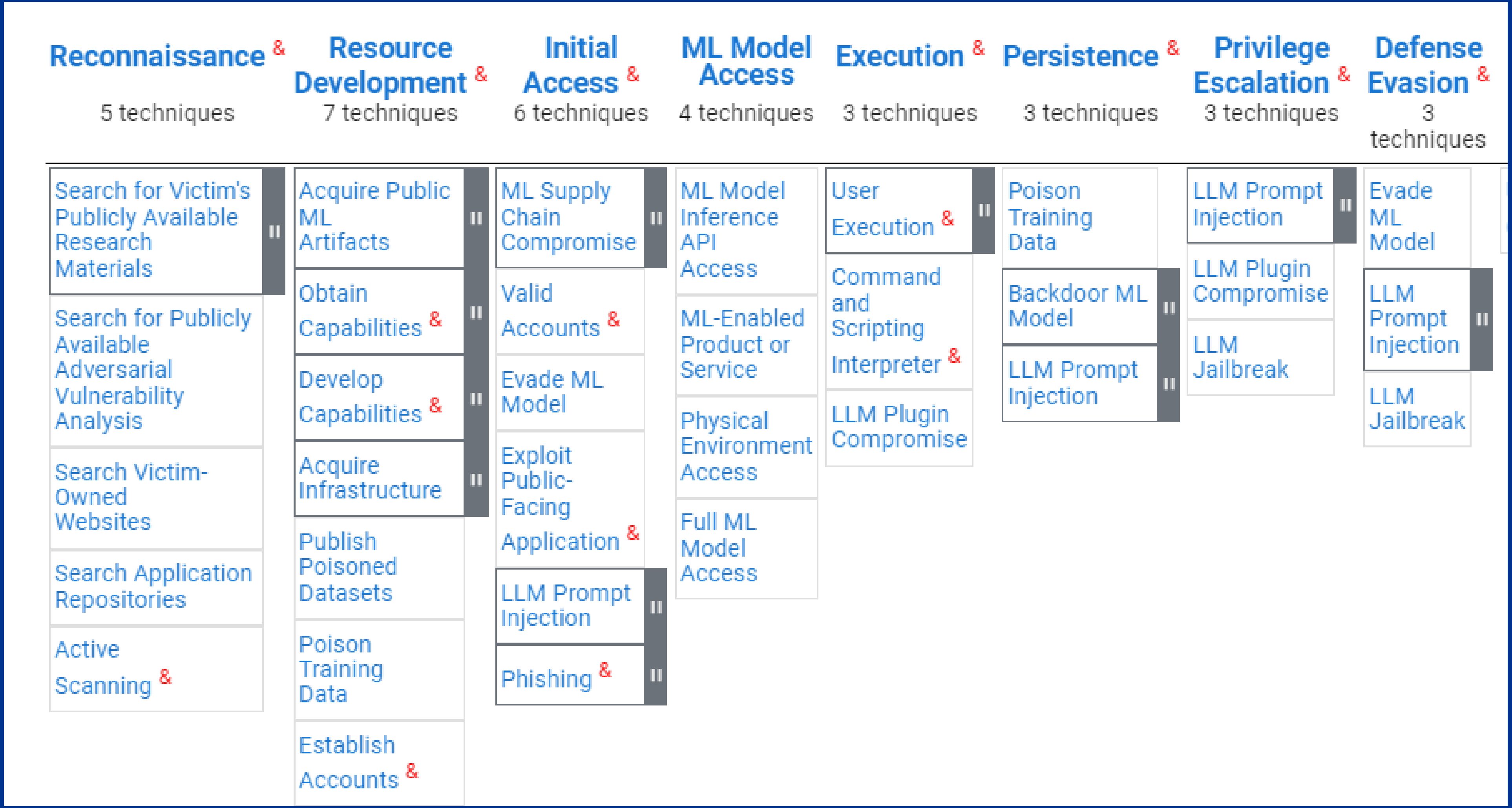
1:6 – 1:20

*In 2021, the average click rate for a phishing campaign was 17,8%. More targeted spear phishing campaigns had an average click rate of 53,2%. About 130 billion spam e-mails are sent worldwide on daily basis.
Junk e-mails represent more than 85% of e-mail communication.

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

MITRE ATLAS

Adversarial Threat Landscape for Artificial-Intelligence Systems



Reconnaissance & 5 techniques	Resource Development & 7 techniques	Initial Access & 6 techniques	ML Model Access 4 techniques	Execution & 3 techniques	Persistence & 3 techniques	Privilege Escalation & 3 techniques	Defense Evasion & 3 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
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
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access				
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection					
	Poison Training Data	Phishing &					
	Establish Accounts &						

MITRE ATLAS™

TTP

Tactics

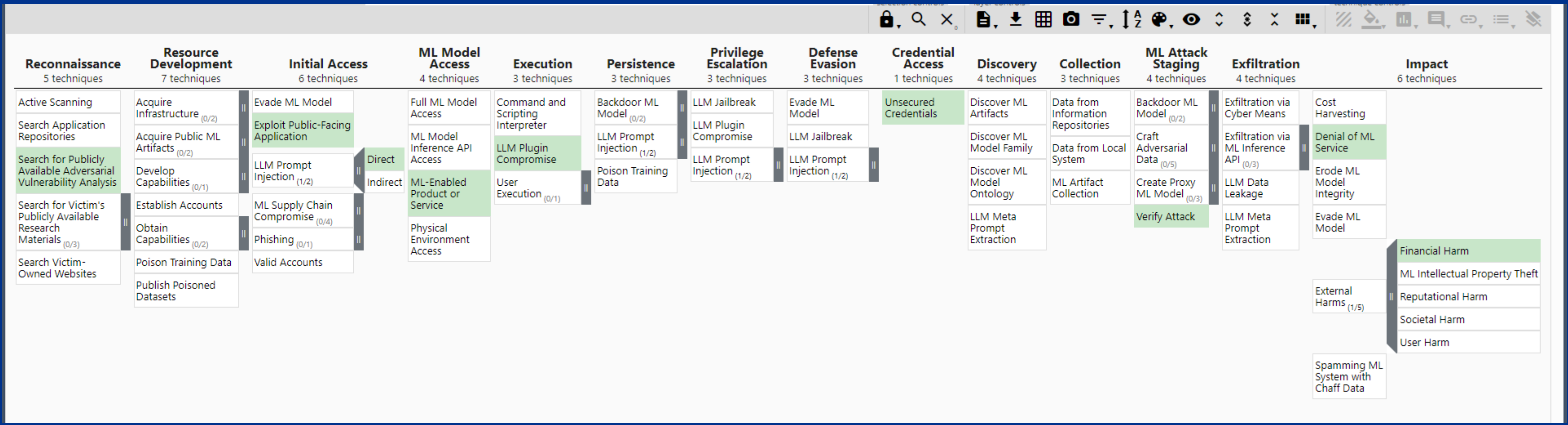
Techniques

Procedures

*ATLAS - Adversarial Threat Landscape for Artificial-Intelligence Systems

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Achieving Code Execution in MathGPT via Prompt Injection



Reconnaissance
5 techniques
- Active Scanning
- Search Application Repositories
- Search for Publicly Available Adversarial Vulnerability Analysis
- Search for Victim's Publicly Available Research Materials (0/3)
- Search Victim-Owned Websites

Resource Development
7 techniques
- Acquire Infrastructure (0/2)
- Acquire Public ML Artifacts (0/2)
- Develop Capabilities (0/1)
- Establish Accounts
- Obtain Capabilities (0/2)
- Poison Training Data
- Publish Poisoned Datasets

Initial Access
6 techniques
- Evade ML Model
- Exploit Public-Facing Application
- LLM Prompt Injection (1/2)
  - Direct
  - Indirect
- ML Supply Chain Compromise (0/4)
- Phishing (0/1)
- Valid Accounts

ML Model Access
4 techniques
- Full ML Model Access
- ML Model Inference API Access
- ML-Enabled Product or Service
- Physical Environment Access

Execution
3 techniques
- Command and Scripting Interpreter
- LLM Plugin Compromise
- User Execution (0/1)

Persistence
3 techniques
- Backdoor ML Model (0/2)
- LLM Prompt Injection (1/2)
- Poison Training Data

Privilege Escalation
3 techniques
- LLM Jailbreak
- LLM Plugin Compromise
- LLM Prompt Injection (1/2)

Defense Evasion
3 techniques
- Evade ML Model
- LLM Jailbreak
- LLM Prompt Injection (1/2)

Credential Access
1 techniques
- Unsecured Credentials

Discovery
4 techniques
- Discover ML Artifacts
- Discover ML Model Family
- Discover ML Model Ontology
- LLM Meta Prompt Extraction

Collection
3 techniques
- Data from Information Repositories
- Data from Local System
- ML Artifact Collection

ML Attack Staging
4 techniques
- Backdoor ML Model (0/2)
- Craft Adversarial Data (0/5)
- Create Proxy ML Model (0/3)
- Verify Attack

Exfiltration
4 techniques
- Exfiltration via Cyber Means
- Exfiltration via ML Inference API (0/3)
- LLM Data Leakage
- LLM Meta Prompt Extraction

Impact
6 techniques
- Cost Harvesting
- Denial of ML Service
- Erode ML Model Integrity
- Evade ML Model
- External Harms (1/5)
  - Financial Harm
  - ML Intellectual Property Theft
  - Reputational Harm
  - Societal Harm
  - User Harm
- Spamming ML System with Chaff Data

MathGPT - interprets user input, writes Python code, executes it and provide the answer

Impact – API key extraction, DoS attack (API limits and budget exhaustion)

Countermeasures - API key rotation and prompt filtering

MITTRE ATT&CK Navigator

*ATT&CK - Adversarial Tactics, Techniques, and Common Knowledge

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Summary

- Regulations and frameworks exist and can be helpful (NIST, MITRE, NSA, CISA, AI Act)
- Specialized AI applications which can be used for almost every imaginable purpose
- Low-code to no-code transformation
- Threat landscape constantly changes
- Adversary tactics will incorporate AI
- AI agents are the future (Auto-GPT, SuperAGI, NExTGPT)

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

AI as a part of cybersecurity defense

Cybersecurity tools and applications

- SIEM, SOAR
- Antivirus, Antimalware
- EDR / NDR / XDR
- E-mail/Spam Proxies
- Physical security – surveillance systems
- UEBA – User and Entity Behavior Analytics
  - added extended identity validation
  - Embedded sensor data, time, location

Microsoft conferences

The Future of Security with AI

Becoming an AI-Powered Organization with Microsoft Copilot

Inside Microsoft AI innovations

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Děkujeme

David Pikálek

dpikalek@kpmg.cz

Principal Consultant

Cyber Security

Lubomír Bartuška

lbartuska@kpmg.cz

Principal Consultant

Cyber Security

kpmg.cz

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.

© 2024 KPMG Česká republika, s.r.o., a Czech limited liability company and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Document Classification: KPMG Public