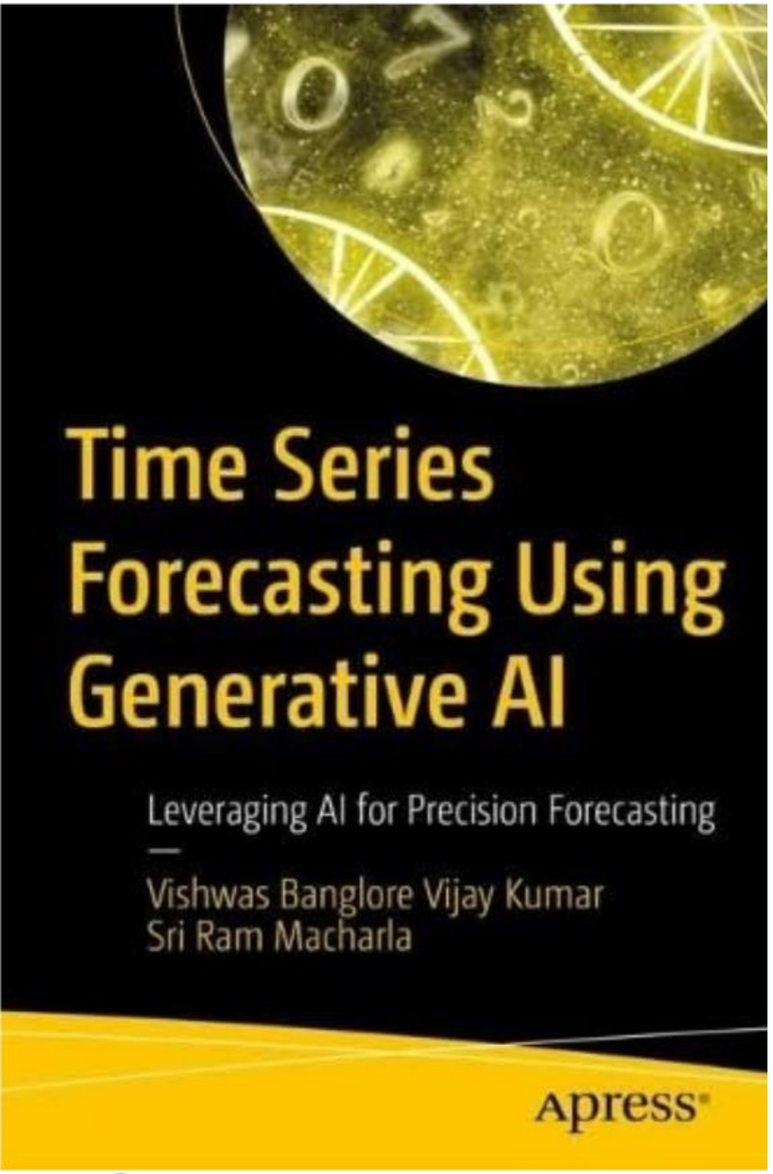
Empowering Organizations Through Innovative Technology

Explore how Involgix partners with organizations to leverage AI, machine learning, and automation for driving efficiency and innovation across various sectors.

Involgix



Sriram Macharla

Exploring speaker's expertise

About the Speaker

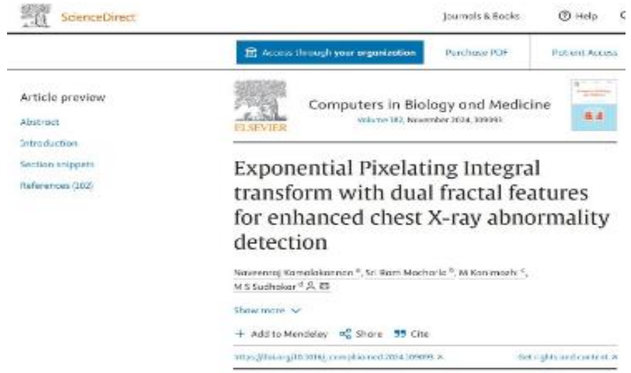
Software architect with focus on AI and Security.

Author

Time Series Forecasting using Generative AI.

AI Research

Publications in major conferences and a journal .



IEEE P1947™ Contributor

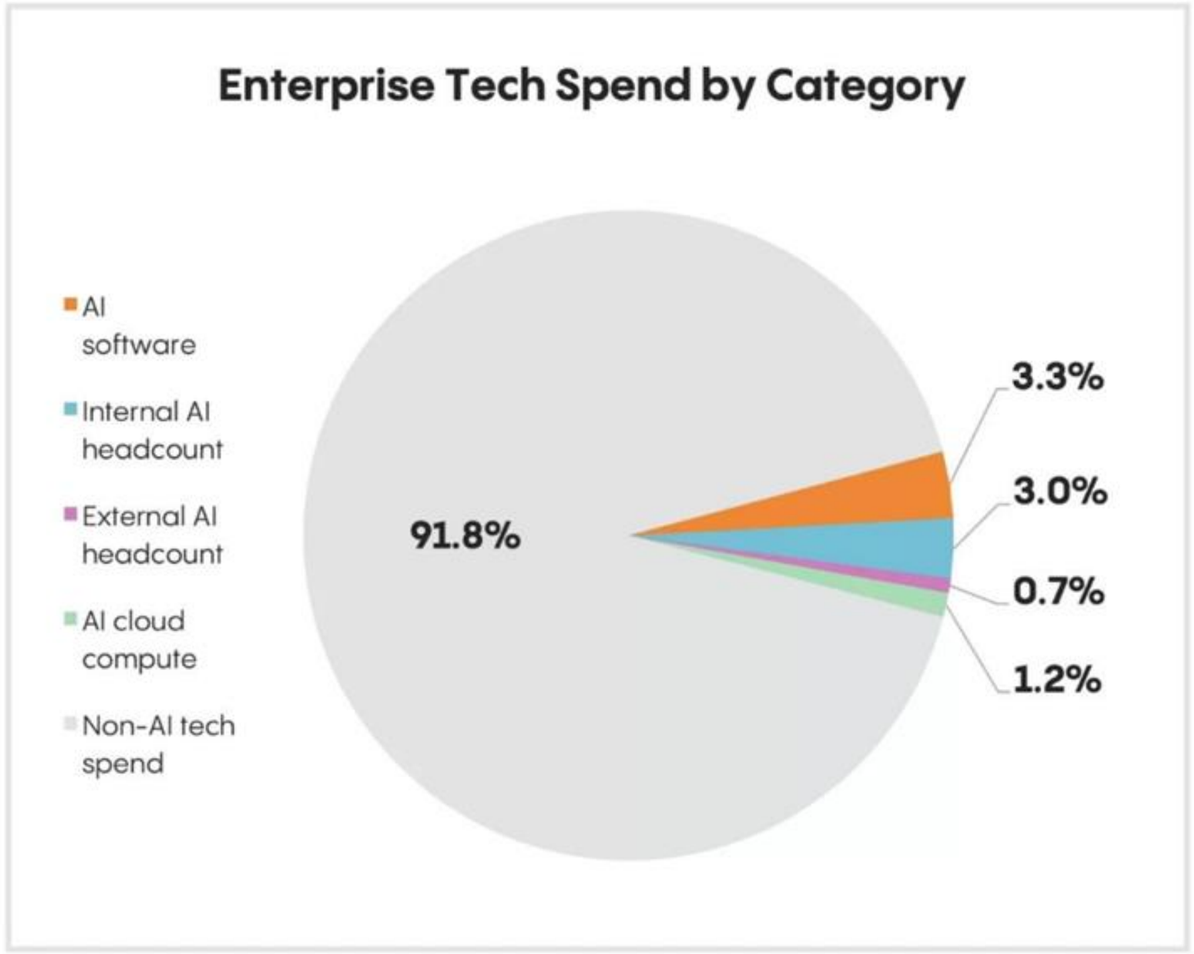
IEEE P1947™, Quantum Cybersecurity Framework standard.

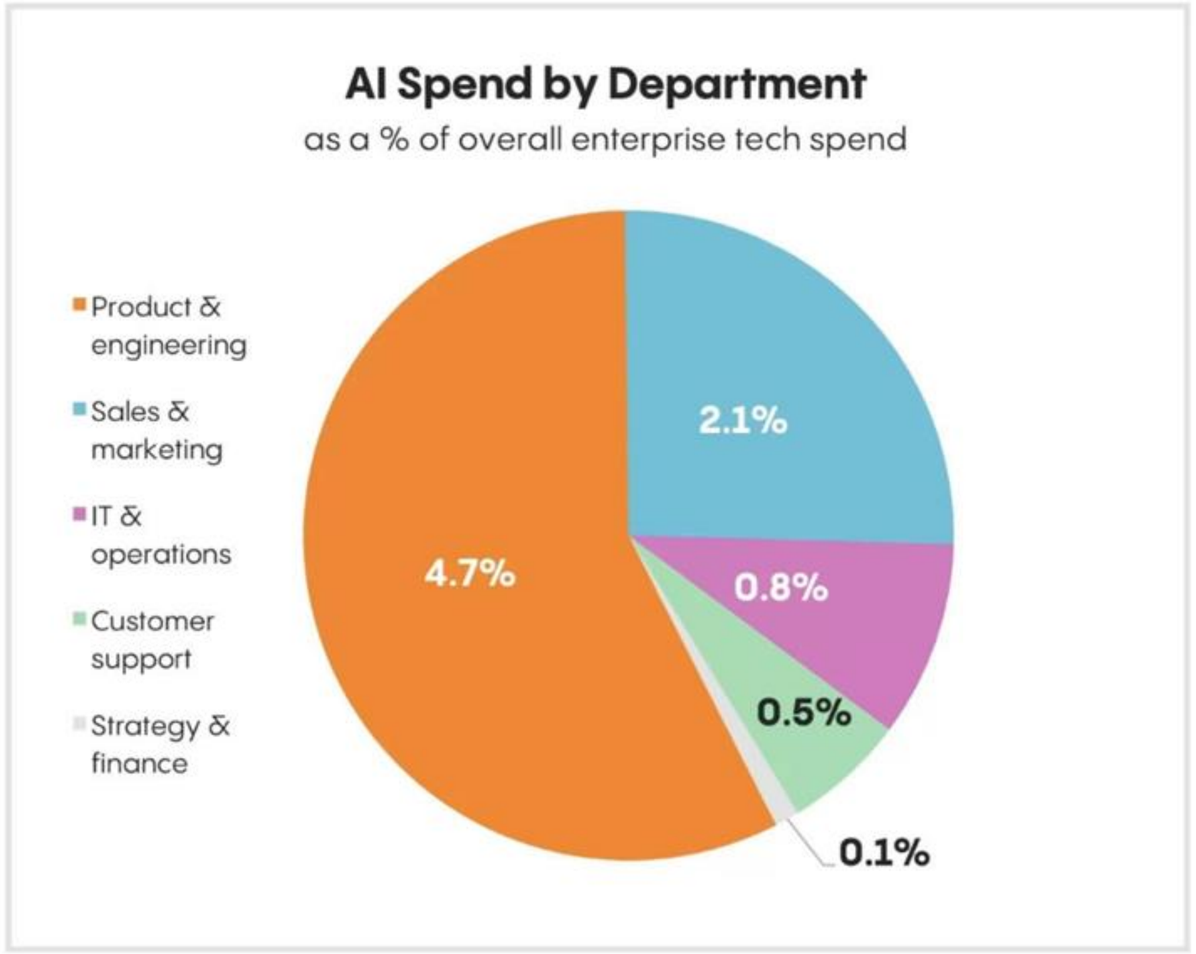
Question-

How many training tokens(%) need to be manipulated to fiddle with the responses from LLM?
Provide your answer as the smallest number(%) you know of?

Technology Spend by Category

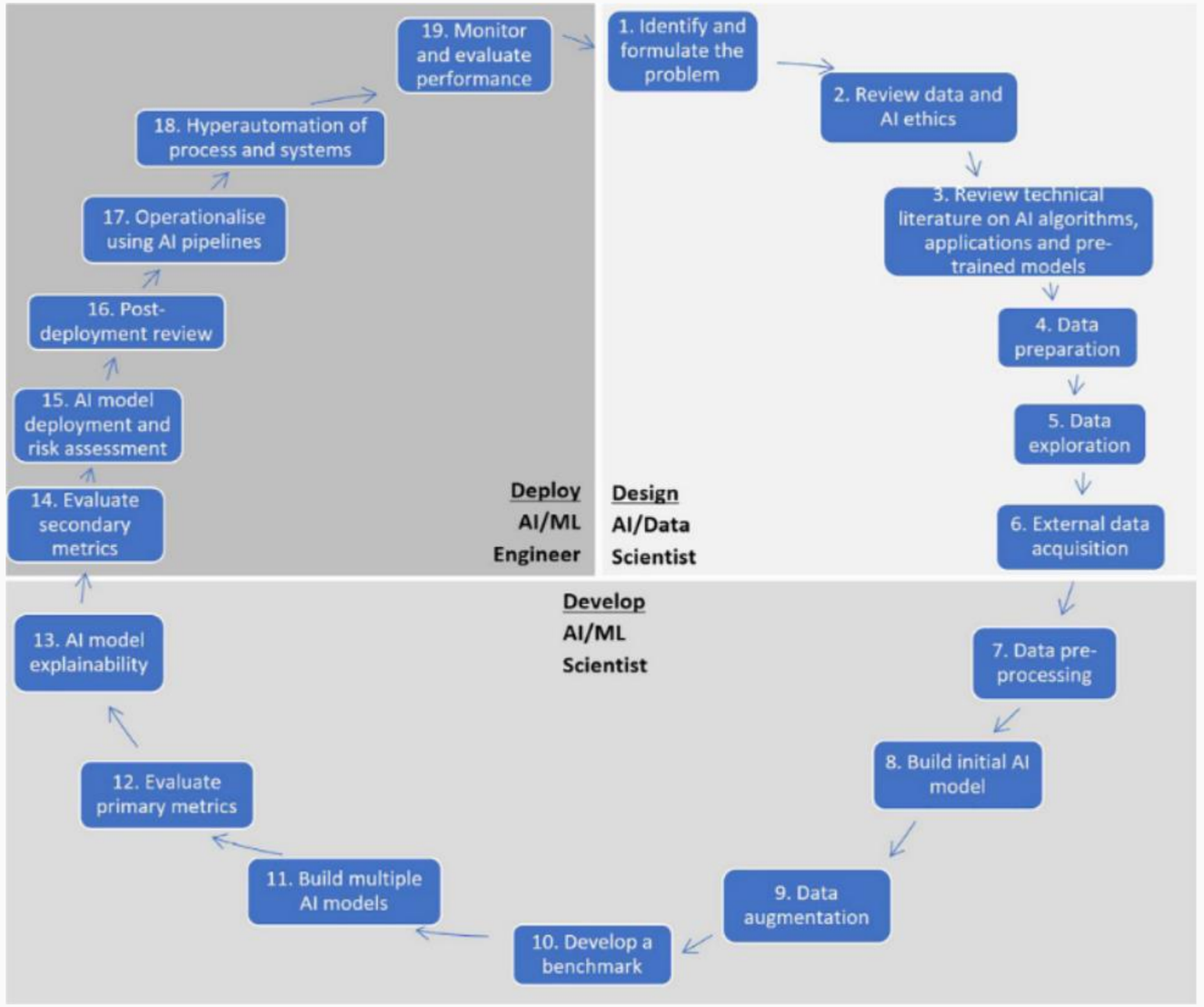
Source: Menlo Ventures 2023





An Artificial Intelligence Life Cycle:From Conception to Production

Image Courtesy: https://doi.org/10.1016/j.patter.2022.100489



Overview of AI Security Layers

Understanding the essential components of AI Security

01

Data Security

Ensures the integrity and confidentiality of data used in AI systems.

02

Model Training Security

Protects AI models during the training phase from adversarial threats.

03

Deployment & Inference Security

Safeguards AI systems during deployment and while making predictions.

04

User Interaction & Ethical Considerations

Focuses on ethical guidelines and secure user interactions with AI.

Understanding AI Security Layers

Exploring threats and solutions in AI security

01

Integration of AI Systems

AI systems are increasingly integrated into critical applications across various industries.

03

Threat Exploration

This presentation will delve into potential threats faced by AI systems at different layers.

02

Comprehensive Security Approach

Security must encompass not only the AI model but the entire AI stack to ensure robustness.

04

Mitigation Strategies

Strategies to mitigate risks and enhance security will be explored for each layer of AI.

Addressing AI Security Challenges

Understanding the key vulnerabilities in AI systems

Data Privacy and Poisoning

Ensuring the protection of user data against unauthorized access and manipulation is crucial.

Adversarial Attacks on Models

Models are vulnerable to inputs designed to deceive, leading to incorrect outputs.

Infrastructure Vulnerabilities

Weaknesses in underlying infrastructure can be exploited, compromising AI systems.

API Security Risks

APIs can be entry points for attacks if not properly secured, exposing sensitive data.

Regulatory and Compliance Concerns

Adhering to laws and regulations is essential for maintaining trust and avoiding penalties.

Exploring The AI Systems Stack

Understanding the Essential Layers of AI Security

01 **Data Security (Input Layer)**

Ensures protection of data entering AI systems to prevent breaches.

02 **Model Security (Processing Layer)**

Safeguards the AI models during processing against malicious attacks.

03 **Infrastructure Security (Hardware & Cloud)**

Secures the physical and cloud infrastructure hosting AI systems.

04 **Application Security (End-User Interaction)**

Protects user interactions with AI applications from vulnerabilities.

05 **Governance & Compliance (Oversight & Regulation)**

Ensures AI operations align with legal and ethical standards.

Data Security Threats and Mitigations

Exploring key threats and effective mitigation strategies

01

Data Poisoning Attacks

Malicious inputs alter datasets, compromising data integrity.

02

Privacy Violations

Non-compliance with regulations like GDPR leads to legal consequences.

03

Unauthorized Access

Breaches result in data leakage and sensitive information exposure.

04

Data Validation & Sanitization

Processes to ensure data integrity and prevent harmful inputs.

05

End-to-End Encryption

Protects data during transfer and storage through encryption protocols.

Model Security Threats and Mitigations

Understanding the Risks and Protection Strategies

Adversarial Attacks

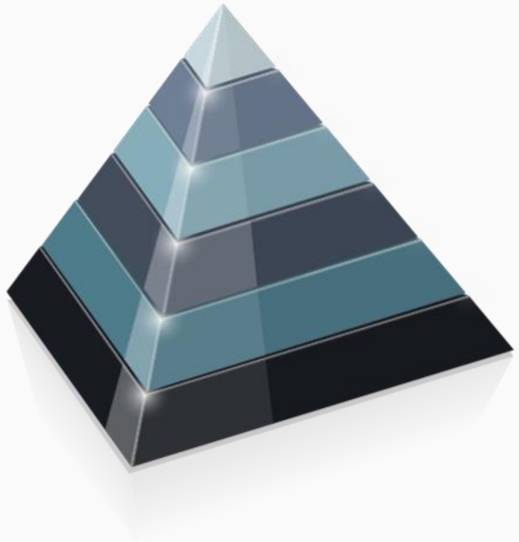
Techniques that manipulate AI models to produce incorrect outputs.

IP Theft

Unauthorized use of intellectual property related to AI models.

Model Watermarking

Embedding unique identifiers in models to protect against misuse.



Model Inversion

Methods used to extract sensitive data from AI models.

Adversarial Training

Training models with adversarial examples to improve robustness.

Secure Federated Learning

Training models across decentralized devices without sharing raw data.

Infrastructure Security Threats Overview

Understanding key threats and effective mitigations

Threats to Infrastructure Security

Identifies major threats impacting infrastructure security.

Cloud API Exploitation

Exploitation of cloud APIs can lead to unauthorized access and data breaches.

Hardware Vulnerabilities

Hardware vulnerabilities, such as GPU attacks, pose significant risks to system integrity.

DDoS Attacks on AI Services

Distributed Denial of Service attacks can cripple AI services, causing downtime and loss.

Mitigation Strategies for Security

Outlines crucial strategies to safeguard infrastructure from threats.

Secure API Authentication

Implementing secure authentication for APIs is vital to prevent unauthorized access.

Regular Hardware Updates

Frequent updates to firmware and hardware help mitigate vulnerabilities.

Zero Trust Architecture

ZTA ensures that no user or system is trusted by default, enhancing security.

Application Security Threats and Mitigations

Exploring critical risks and effective mitigations

AI Bias and Fairness Issues

Addressing biases in AI algorithms is crucial to ensure fairness and prevent discrimination.

Prompt Injection Attacks

These GenAI risks involve manipulating input prompts to exploit vulnerabilities in AI systems.

Model Drift Impact

Model drift can lead to inaccurate predictions over time, requiring regular updates and evaluations.

Explainable AI (XAI)

Implementing XAI enhances transparency, making AI decision processes clearer to users.

Continuous Monitoring

Ongoing monitoring and logging of AI systems help identify and mitigate emerging threats effectively.

Human-in-the-Loop Oversight

Incorporating human oversight for critical AI decisions mitigates risks and enhances accountability.

Governance & Compliance Threats and Mitigations

Understanding threats to governance and compliance in AI

01 **Regulatory Non-Compliance Risks**

Organizations face significant risks by failing to comply with regulations, which can result in fines and penalties.

02 **Lack of Auditability & Explainability**

A lack of ability to audit AI systems can hinder transparency and accountability in decision-making processes.

03 **Reputational Damage from AI Failures**

Failures of AI systems can lead to severe reputational damage, impacting customer trust and brand loyalty.

04 **AI Risk Assessments & Security Audits**

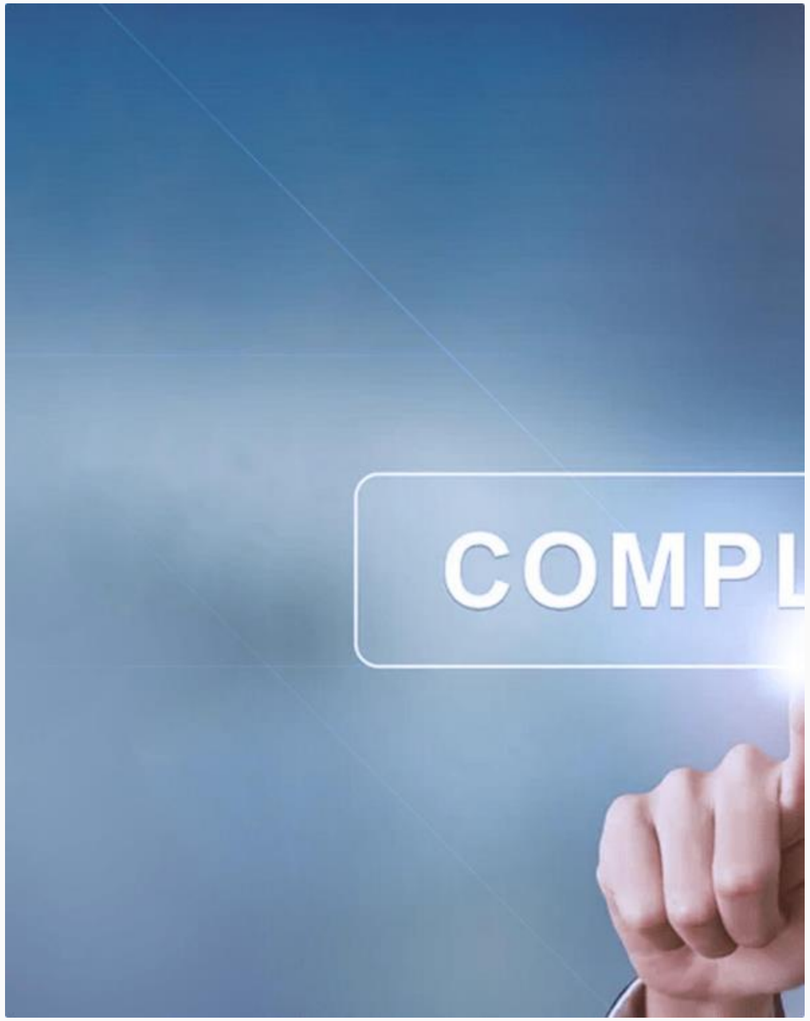
Conducting thorough risk assessments and security audits is vital to identify vulnerabilities and protect data.

05 **Alignment with Regulations**

Aligning AI operations with GDPR, AI Act, and NIST frameworks ensures compliance and mitigates legal risks.

06 **Ethical AI Policies**

Implementing ethical AI policies fosters responsible use of technology and promotes stakeholder trust.



Best practices

Implementing robust strategies for AI security

Multi-layered approach is essential

AI security necessitates a comprehensive strategy, addressing various potential vulnerabilities.

Proactive threat mitigation

Anticipating threats and implementing solutions ensures that AI systems remain reliable and functional.

Investing in security by design

Building AI systems with security integrated from the start fosters greater trust among users and stakeholders.

Evolution of AI Architecture: Traditional ML to Generative AI

Traditional ML

Data Pre-Processing
Cleaning and preparing data for analysis

Feature Engineering
Extracting important features from data

Training & Tuning
Training models on data and adjusting parameters for optimal performance

Deployment & Monitoring
Implementing models in real-world applications and monitoring their performance

Generative AI

Data Pre-Processing
Cleaning and preparing data for analysis

Prompt Engineering/Fine Tuning
Designing effective prompts to guide AI in generating desired outputs

Foundational/Fine-Tuned LLM
Using foundation and Fine-tuned language models for sophisticated content generation

Deployment & Monitoring
Implementing models in real-world applications and monitoring their performance

Tech Stack for Traditional ML

- **ML Frameworks:** Keras, Theano
- **ML API's & SDK:** IBM Watson
- **Database:** SQL Server, Oracle
- **ML Ops:** Docker, Jenkins

Tech Stack for Generative AI

- **Gen AI Orchestration:** Langchain, llamaindex
- **LLM Models:** OpenAI, Anthropic
- **Vector Database:** Pinecone, Weaviate
- **LLM Ops:** Prompt Layer, Helicone

aim research

Ensuring Security in Generative AI Stack

Mitigating Risks in Generative AI Technologies

Identifying Vulnerabilities

Potential security gaps within the generative AI stack.

Data Privacy Concerns

Addressing the risks of sensitive data exposure during AI training.

Model Integrity Assurance

Ensuring that AI models are not tampered with to maintain their accuracy.

Adopting Secure Coding Practices

Implementing best practices for secure software development in AI.

Regular Security Audits

Conducting frequent audits to identify and rectify security issues.

User Awareness and Training

Educating users on security risks associated with generative AI.

Compliance with Regulations

Ensuring adherence to laws and standards governing AI systems.

Incident Response Planning

Developing a robust plan to address security breaches effectively.

Understanding Generative AI Risks

Exploring the implications of AI-generated content

01

What is Generative AI?

Generative AI refers to algorithms that create content, including text, images, and audio.

02

Why Security Matters?

Security is crucial because AI systems can produce harmful or misleading content if not properly managed.

03

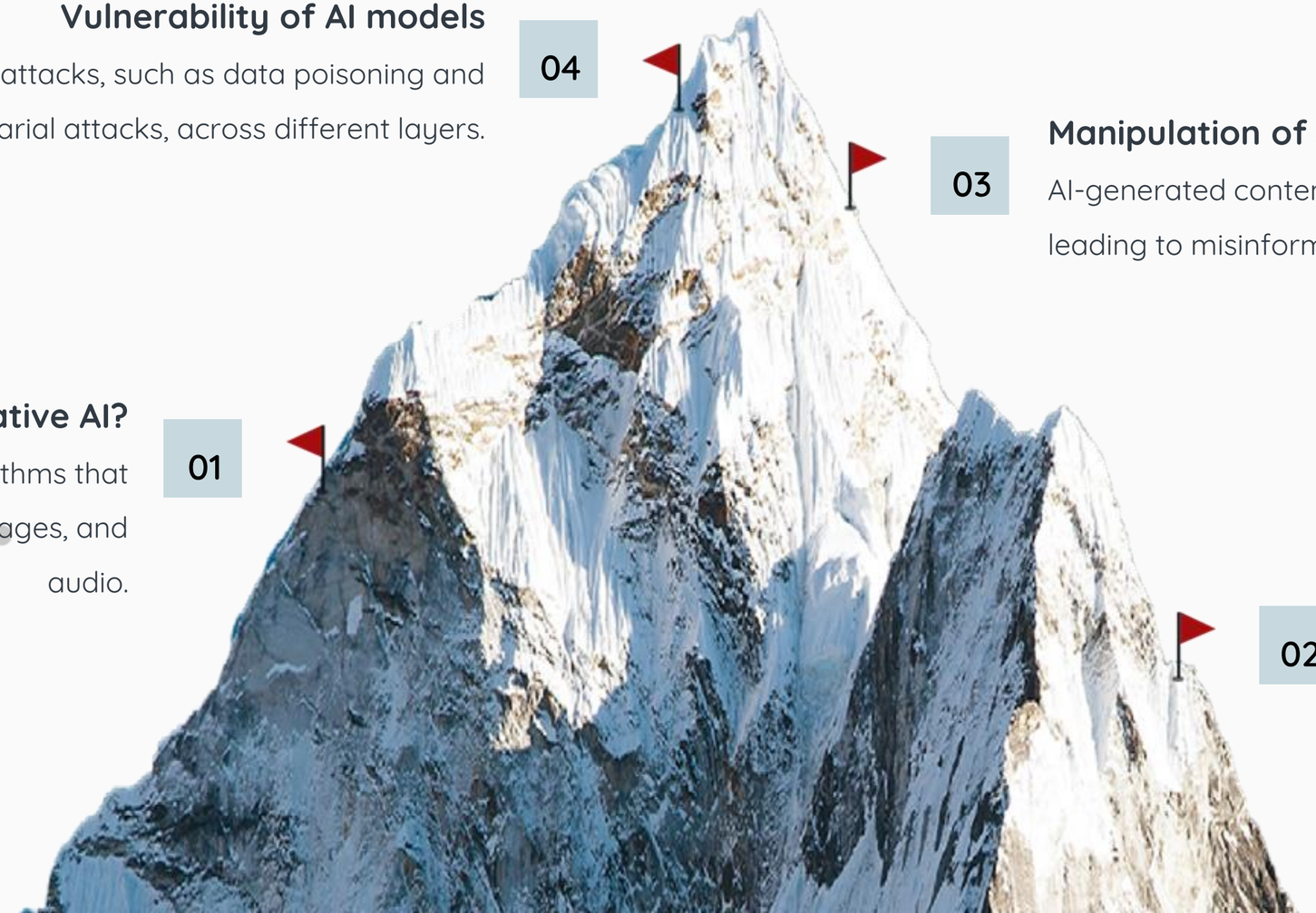
Manipulation of AI-generated content

AI-generated content can be easily manipulated, leading to misinformation and unethical uses.

04

Vulnerability of AI models

AI models face various attacks, such as data poisoning and adversarial attacks, across different layers.



Understanding Data Layer Security

Challenges and Mitigation in Data Security

Data Poisoning Attacks

Malicious actors can corrupt training data, leading to ineffective models.

Privacy Risks (PII Leakage)

Exposure of Personally Identifiable Information can violate user privacy.

Bias in Training Data

Inherent biases can skew model predictions and perpetuate discrimination.

Secure Data Pipelines

Implementing robust security measures throughout data processing can mitigate risks.

Differential Privacy & Federated Learning

These techniques enhance user privacy while maintaining data utility.

Data Auditing & Bias Detection

Regular audits can identify biases and ensure data integrity over time.

Model Training & Pipeline Security

Addressing Security Challenges in AI Model Training

01 **Model Poisoning & Backdoor Attacks**

These attacks compromise model integrity by injecting malicious data during training.

02 **Adversarial Manipulations**

Adversaries can alter input data to deceive the model, impacting its performance.

03 **Supply Chain Vulnerabilities**

Weaknesses in the software supply chain can lead to exploited components, threatening security.

04 **Adversarial Training**

This technique involves training models with adversarial examples to enhance robustness.

05 **Homomorphic Encryption**

This method allows computations on encrypted data, protecting sensitive information.

06 **Secure Software Supply Chains**

Implementing security measures in the software supply chain to mitigate risks.

Deployment & Inference Security

Addressing Security Risks in AI Deployment

01 **Model Theft**

API scraping poses a risk of model theft, compromising proprietary algorithms.

02 **Data Leakage**

Responses from APIs can unintentionally leak sensitive data, risking privacy.

03 **API Vulnerabilities**

APIs may have security vulnerabilities that can be exploited by attackers.

04 **Rate Limiting**

Implementing API rate limiting can prevent abuse and unauthorized access.

05 **Access Control**

Restricting access helps safeguard against unauthorized users interacting with APIs.

06 **Watermarking**

Watermarking AI-generated content can help trace origin and mitigate misuse.

07 **Red Teaming**

Conducting AI red teaming exposes vulnerabilities through simulated attacks.

08 **Stress Testing**

Stress testing APIs can identify weaknesses under high load conditions.

User Interaction & Ethical Risks

Addressing the Ethical Risks of AI in User Interactions

Challenges of AI in User Interaction

AI-generated content poses risks like social engineering and deepfake attacks.

Social Engineering Threats

AI can create deceptive content that misleads users into compromising information.

Deepfake Impersonation Attacks

Deepfakes can be used to impersonate individuals, leading to identity fraud.

Ethical Concerns with AI Media

The use of AI in media raises questions about authenticity and accountability.

AI Content Detection Solutions

Implementing AI detection tools can help identify and verify content authenticity.

Importance of Transparent AI Disclosures

Clear disclosures about AI-generated content can help users understand its origin.

Enhancing Public Awareness

Educating the public about AI risks is crucial for informed interactions.

Case Studies of AI Incidents

Examining notable incidents in AI technology

Microsoft Tay Chatbot Poisoning

A significant incident where users manipulated Tay to make offensive comments.

Tesla Autopilot Attacks

Adversarial attacks that misled Tesla's Autopilot system, risking safety.

GPT-3 Model Extraction Attempts

Attempts to extract and replicate the powerful GPT-3 model, raising concerns about intellectual property.

Deepfake Financial Scams

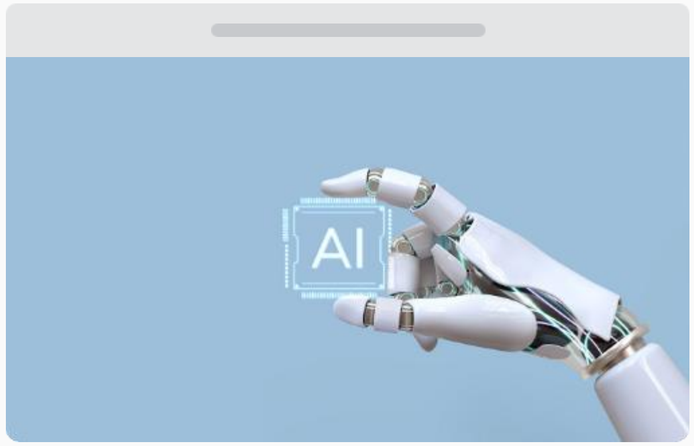
Use of deepfake technology in scams, leading to significant financial fraud.

Trends in AI Security and Threats

Navigating the Future of AI Regulations and Solutions

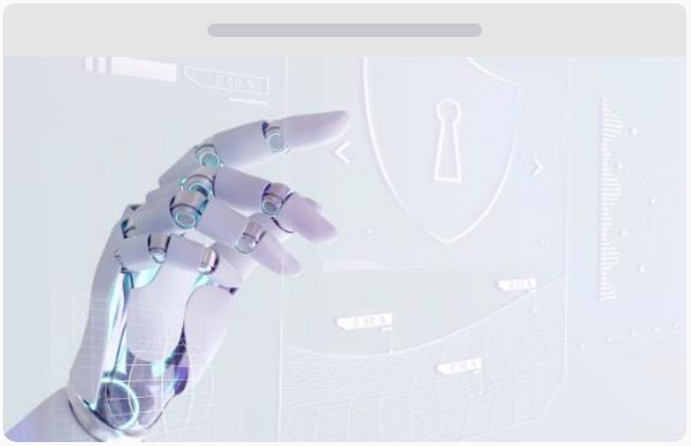
Emerging AI Security Regulations

The EU AI Act and NIST Framework aim to establish guidelines for AI safety and accountability.



Aligning AI with Human Values

Ensuring AI systems are developed and deployed with a focus on aligning with human ethical standards.



Next-Gen AI-Driven Security Solutions

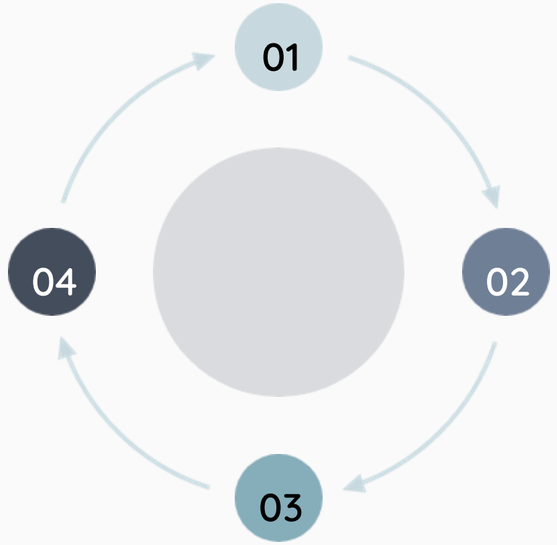
Innovative security solutions that leverage AI to enhance cybersecurity measures and threat detection.

Essential AI Risk Management Resources

Explore key resources for AI risk management and security insights

NIST AI Risk Management Framework

A comprehensive guide for managing AI risks effectively.



MITRE ATLAS Overview

Explores adversarial threats and vulnerabilities in AI systems.

Google AI Security Insights

Latest findings and trends in AI security from Google.

OpenAI Security Blogs

Insights and updates on AI security practices and research.

Understanding the AI Security Landscape

The necessity of adapting to AI advancements

Evolving Threat Landscape

Acknowledging the dynamic nature of AI threats is crucial for proactive security measures.

Continuous Security Updates

As technology advances, security practices must be continuously updated to mitigate risks.

Importance of AI Security Frameworks

Exploring frameworks that enhance AI security

Understanding GDPR

The General Data Protection Regulation (GDPR) sets standards for data privacy and protection.

Overview of the AI Act

The AI Act establishes a legal framework for the ethical and safe use of AI technologies.

NIST Guidelines

NIST provides a comprehensive framework for improving the security of AI systems.

Framework Integration

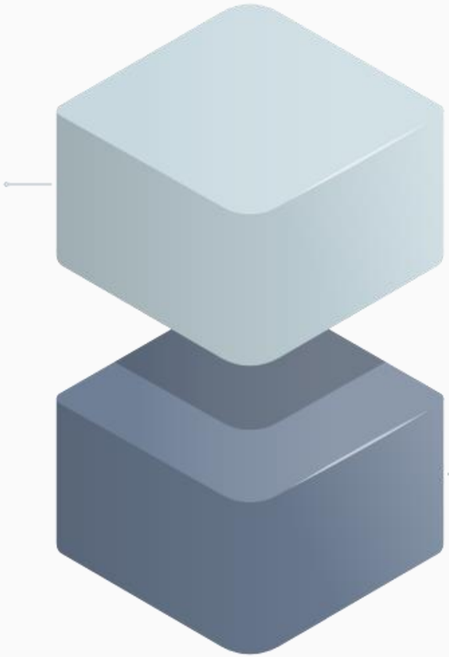
These frameworks guide organizations in implementing secure AI systems effectively.

Innovations in AI Security Ahead

Examining the evolution and future innovations in AI security

Future Directions in AI Security

Exploring emerging trends and technologies shaping AI security landscapes is critical for proactive measures.



AI's Role in Security Solutions

AI is pivotal in crafting next-gen security solutions, enhancing threat detection and response capabilities.

Conclusion & Q/A

Key insights and discussions on AI security measures

Open floor for audience questions

Invite participants to engage with questions and clarifications on the topic.

Future research directions

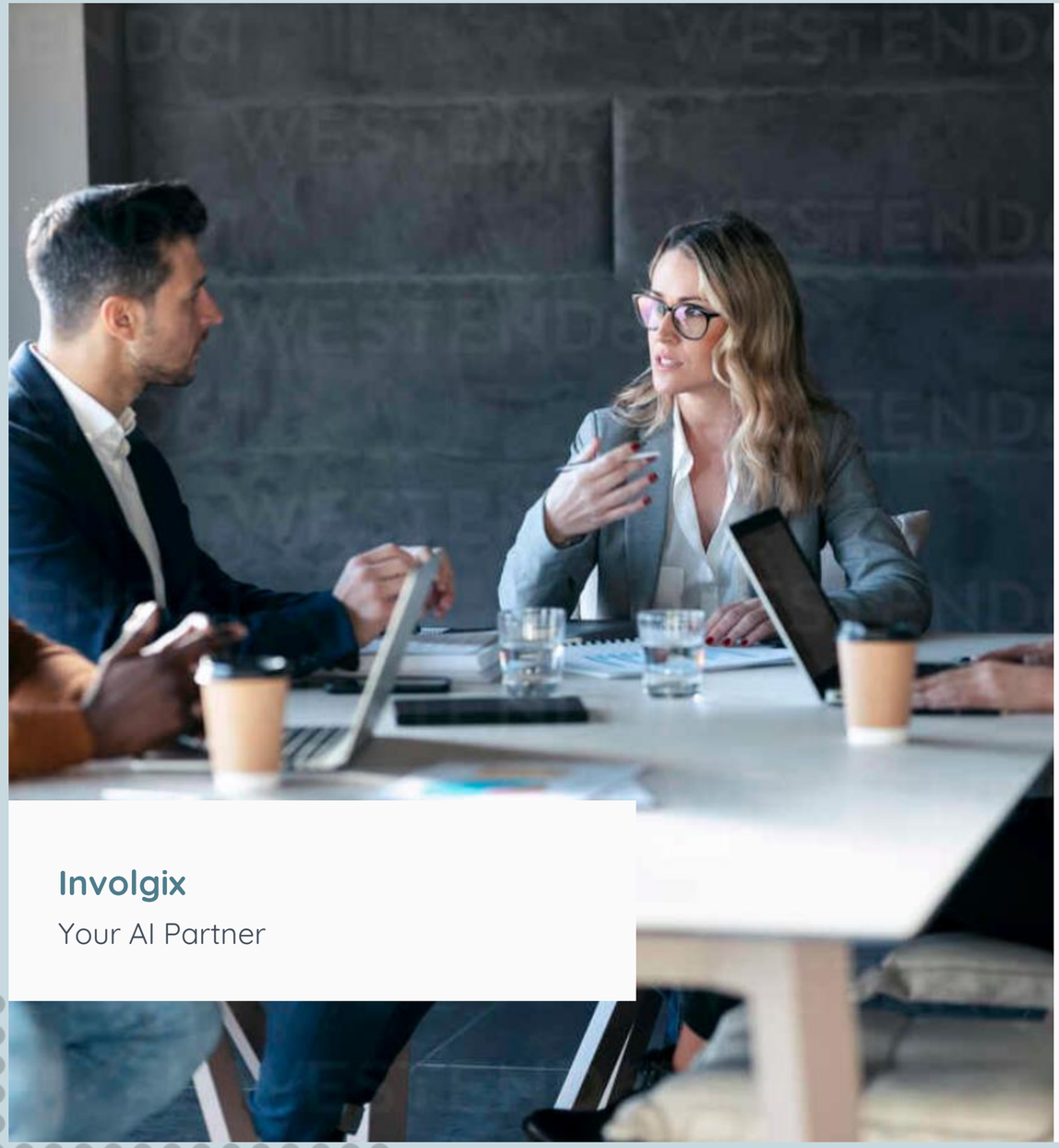
Discuss potential areas for further exploration in AI security advancements.

Recap of key AI security concerns

Summarize the main issues surrounding AI security, highlighting vulnerabilities.

Actionable steps for AI security

Outline practical measures organizations can implement to enhance AI security.



Involgix

Your AI Partner

Thank You

sriram.macharla@involgix.com

Website & Linkedin

www.involgix.com

www.linkedin.com/involgix/

Location

700 Market St ste 203,

Cedar Park, TX 78613