Altimetrik

Altimetrik AI Security

WHITEPAPER

Executive Summary and Challenges

This comprehensive datasheet provides an in-depth overview of the AI security landscape, detailing current threats, vulnerabilities, incidents, and the complex challenges posed by adversarial AI techniques. It examines the top vulnerabilities specific to machine learning (ML) and large language models (LLMs), presenting real-world case studies that highlight the pressing need for robust security measures.

We cover the significance of governance, compliance, and regulation in AI, analyzing key legislative and standardization efforts globally. This includes the EU Artificial Intelligence Act, the impact of GDPR on AI, the Blueprint for an AI Bill of Rights in the US, and various standards and frameworks such as ISO/IEC 42001 and the NIST AI Risk Management Framework. These sections emphasize the evolving landscape of AI regulation and the importance of adhering to these standards for ethical and secure AI deployment.

We also showcase the comprehensive suite of services encompassing automation, AI OPS integration, security self-scanning, and AI-enhanced threat detection. The datasheet explains Altimetrik's proprietary solutions, including the Empulse GenAI framework designed to transform data into actionable insights across industries, and its emphasis on navigating GenAI challenges.

The technical depth of the datasheet extends to AI/ML architecture risk analysis, threat modeling, AI-driven attack maps, and comprehensive AI framework assessments. It highlights innovative approaches to identifying, classifying, and mitigating security issues, such as AI Red Teaming, LLM assessments, and the critical role of AI in enhancing incident response capabilities.

Furthermore, the document explores advancements in security automation and the mitigation of risks associated with disinformation campaigns and supply chain vulnerabilities. Altimetrik's methodologies for mapping the threat landscape, enhancing return on investment (ROI), and developing robust AI security controls are detailed, providing insights into the firm's holistic approach to safeguarding AI environments.

In summary, this comprehensive datasheet offers an in-depth exploration of the challenges and solutions in AI security, showcasing Altimetrik's expertise and innovative approaches to protecting AI infrastructures against threats. It serves as a vital resource for organizations seeking to understand and implement effective AI security strategies, ensuring the ethical, compliant, and secure utilization of AI technologies.

Glossary

Artificial Intelligence (AI): encompasses all computer science fields enabling machines to perform tasks requiring human intelligence, with machine learning and generative AI as subcategories.

Machine Learning (ML): a subset of AI, entails creating algorithms that can learn from data and make predictions or decisions based on that data.

Deep Learning: Deep learning is a branch of artificial intelligence that mimics the workings of the human brain to process data and make predictions.

Generative AI (GenAI): refers to a category of artificial intelligence (AI) algorithms that generate new outputs based on the data they have been trained on. These outputs can include text, images, audio, and more. Generative AI models become more sophisticated with the more data they receive, resulting in outputs that are increasingly convincing and human-like

Large Language Model (LLM): is an AI model that processes and generates human-like text. In the context of artificial intelligence, a "model" is a system trained to make predictions based on input data. LLMs are specifically trained on large natural language datasets, hence the name "large language models."

Retrieval-Augmented Generation (RAG): is an AI framework that retrieves facts from an external knowledge base to ground large language models (LLMs) on the most accurate, up-to-date information and to provide insight into LLMs' generative process.

Adversarial Machine Learning (AML): Refers to a technique used in machine learning to deceive or misguide a model with malicious input. It aims to trick machine learning models by providing deceptive input, including the generation and detection of adversarial examples designed to deceive classifiers. This technique is commonly used to execute attacks or cause malfunctions in machine learning systems, and it has been extensively explored in areas such as image classification and spam detection.

Targeted poisoning: Poisoning attacks against machine learning that change the prediction on a small number of targeted samples.

Backdoor poisoning: Refers to a technique where an attacker intentionally injects a specific pattern or trigger (known as a backdoor) into the training data of a machine learning model. This backdoor is designed in such a way that when the trained model encounters the specific trigger pattern during inference (making predictions), it behaves in a way desired by the attacker, rather than its intended behavior.

Model poisoning: Model poisoning refers to poisoning attacks in which the model parameters are under the control of the adversary.

Data privacy attacks: Attacks against machine learning models to extract sensitive information about training data

Adversarial examples: Are modified testing samples that induce misclassification of a machine learning model at deployment time

Generative Adversarial Networks (GAN): is a machine learning framework where two neural networks engage in a zero-sum game, learning to generate new data with the same statistics as the training set

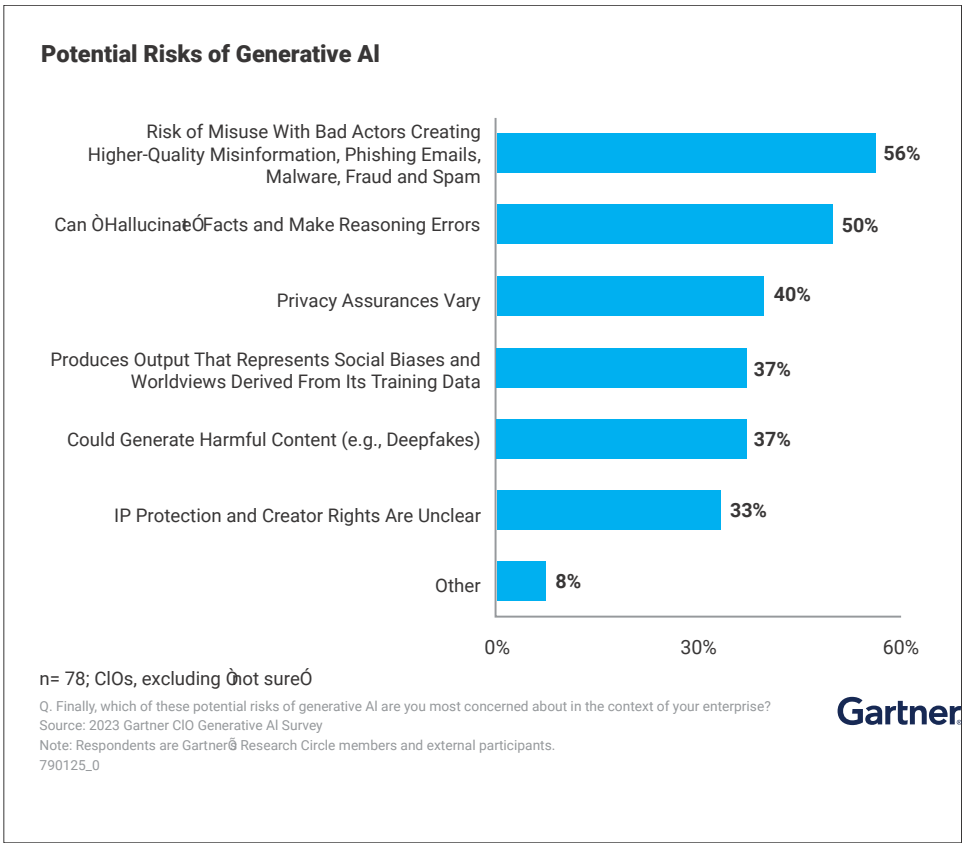
Optical Character Recognition (OCR): Technology used to convert different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data. OCR software analyzes the shapes of characters in scanned images or documents and translates them into machine-readable text.

The AI Threat Landscape

As AI technologies become increasingly sophisticated and pervasive, so do the methods and motivations of malicious actors seeking to exploit them – covered in the attacks and vulnerabilities section. The following are examples of how AI is changing the threat landscape:Regulatory and compliance challenges - The U.S. emphasizes safety, security, and ethics, Europe balances innovation with risk, and China prioritizes advancement over regulation

- **Data Poisoning:** Increasing instances of data poisoning where malicious actors inject misleading or biased data into AI/ML training datasets, compromising model integrity.
- Increased sophistication of social engineering attacks with AI-enhanced phishing
- Impersonation attacks and the proliferation of deep fakes
- The exploitation of bias in AI models
- AI Duality – the weaponization of AI for disinformation campaigns, the automation of attacks and adversarial machine learning
- Supply chain attacks on AI dependencies rapid exploitation of vulnerabilities
- Development of complex AI-driven malware code and malware mutation
- **Ransomware evolution:** AI enhances ransomware's ability to evade detection, customize payloads, and surgically target high-value data assets. Threat actors leverage AI to prioritize the exfiltration of critical information like trade secrets before broadly encrypting files, making attacks more impactful.
- Model theft and LLM model exploitation
- Rise of digital surveillance and loss of privacy

Adversaries increasingly leverage LLM and Generative AI tools to refine traditional methods of attacking organizations, individuals, and government systems. LLM facilitates the enhancement of techniques for new crafting malware, potentially embedding novel zero-day vulnerabilities or evading detection. It also enables the creation of convincing deep fakes for social engineering ploys and the development of innovative hacking capabilities. Criminal use of AI technology will require specific responses and dedicated solutions for an organization's defense and resilience. Organizations also face the threat of not utilizing LLM capabilities, leading to a competitive disadvantage, outdated market perception, limited scalability of personalized communications, innovation stagnation, operational inefficiencies, higher risk of human error, and inefficient allocation of human resources.



Attacks and Vulnerabilities

As artificial intelligence, machine learning, and large language models continue to advance and affect various aspects of our lives, the impact of attacks and vulnerabilities on these technologies becomes an increasingly critical concern. The emergence of sophisticated cyber threats and the potential for malicious exploitation pose significant challenges to the security and integrity of AI, ML, and LLM systems. This section discusses the implications of attacks and vulnerabilities on these technologies, highlighting the need for defenses, proactive research, and collaborative efforts to safeguard against potential risks.

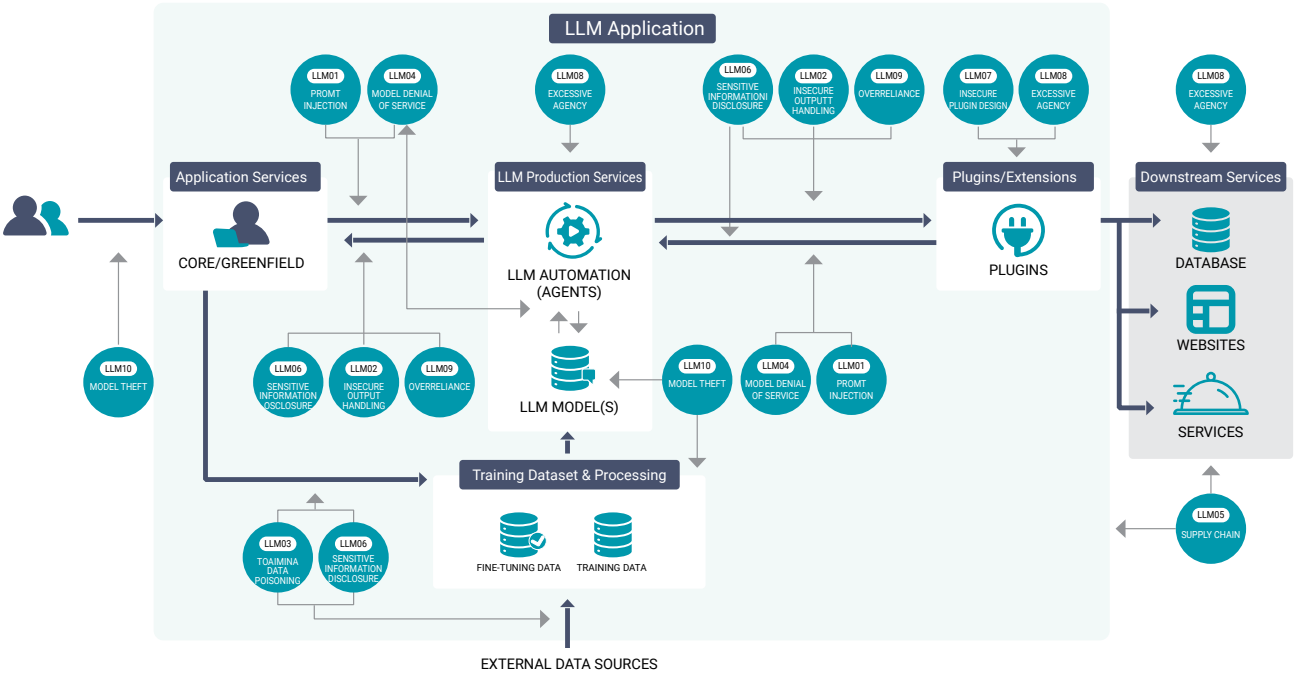
To provide insight into the next few sections, we will first cover the vulnerabilities affecting ML and LLMs in 2024:

Top 10 ML Vulnerabilities

- **Input Manipulation Attack:** A broad category that encompasses Adversarial Attacks, a form of attack wherein a malicious actor intentionally modifies input data to deceive the model.
- **Data Poisoning Attack:** Data poisoning attacks occur when a perpetrator manipulates the training data to induce the model to exhibit undesirable behavior.
- **Model Inversion Attack:** Model inversion attacks occur when an attacker reverse-engineers the model to extract information from it.
- **Membership Inference Attack:** Membership inference attacks occur when an attacker alters the model's training data to induce behavior that exposes sensitive information.
- **Model Theft:** Model theft attacks occur when an attacker obtains access to the model's parameters.
- **AI Supply Chain Attacks:** AI Supply Chain Attacks happen when an attacker alters or substitutes a machine learning library or model, including its associated data, used by a system.
- **Transfer Learning Attack:** Transfer learning attacks occur when an attacker trains a model on one task and then fine-tunes it for another task to induce undesirable behavior.
- **Model Skewing:** Model skewing attacks happen when an attacker manipulates the training data distribution to induce undesirable behavior in the model.
- **Output Integrity Attack:** In an Output Integrity Attack, an attacker seeks to alter the machine learning model's output to change its behavior or harm the system it serves.
- **Model Poisoning:** Model poisoning attacks happen when an attacker manipulates the model's parameters to induce undesirable behavior.

Top 10 LLM Vulnerabilities

- **Prompt Injections:** This involves manipulating a large language model (LLM) with clever inputs to trigger unintended actions. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.
- **Insecure Output Handling:** This vulnerability arises when an LLM output is accepted without scrutiny, exposing backend systems. Misuse can lead to severe consequences such as cross-site scripting (XSS), client-side request forgery (CSRF), server-side request forgery (SSRF), privilege escalation, or remote code execution.
- **Training Data Poisoning:** This happens when LLM training data is tampered with, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, and books.
- **Denial of Service:** Attackers initiate resource-heavy operations on LLMs, resulting in service degradation or high costs. The vulnerability is amplified by the resource-intensive nature of LLMs and the unpredictability of user inputs.
- **Supply Chain Vulnerabilities:** Vulnerable components or services can compromise the LLM application lifecycle, leading to security attacks. The use of third-party datasets, pre-trained models, and plugins can introduce vulnerabilities.
- **Sensitive Information Disclosure:** LLMs may inadvertently expose confidential data in their responses, resulting in unauthorized data access, privacy violations, data privacy attacks, and security breaches. Implementing data sanitization and strict user policies is crucial to mitigate this risk.
- **Insecure Plugin Design:** LLM plugins with insecure inputs and inadequate access control are easier to exploit and can lead to consequences such as remote code execution.
- **Excessive Agency:** LLM-based systems may take actions leading to unintended consequences due to excessive functionality, permissions, or autonomy granted to them.
- **Overreliance:** Over-reliance on LLMs without oversight can lead to misinformation, miscommunication, legal issues, and security vulnerabilities due to the generation of incorrect or inappropriate content.
- **Model Theft:** This entails unauthorized access, copying, or exfiltration of proprietary LLM models, resulting in economic losses, compromised competitive advantage, and potential access to sensitive information.



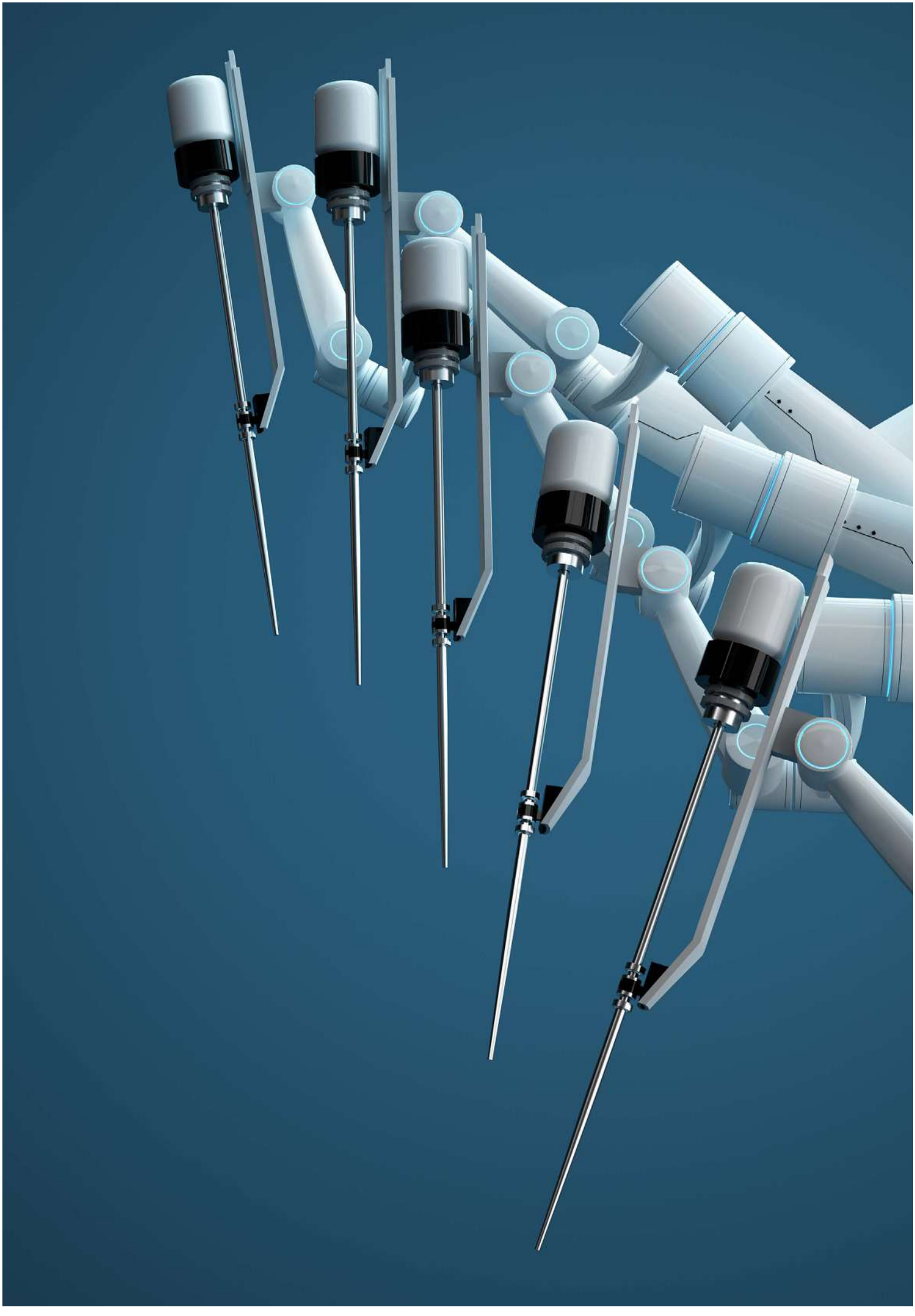
The screenshot above is a high-level overview of the Top 10 LLM vulnerabilities visualized.

AI Security Incidents

The following is a list of real-world examples of abuses and attacks on AI, ML, and LLM assets:

- **Hugging Face ML silent backdoors (2024):** Recent reports have highlighted the emergence of malicious tactics targeting data scientists and AI/ML engineers through Hugging Face ML models with silent backdoors. At least 100 instances of malicious AI ML models were found on the Hugging Face platform, some of which can execute code on the victim's machine, giving attackers a persistent backdoor.
- **ComPromptMized Morris II Worm (2024):** A zero-click worm was created by Cornell Tech students to target GenAI models such as Gemini Pro, ChatGPT 4.0, and LLaVA using adversarial self-replicated prompts. This triggers the GenAI to perform malicious activities such as the distribution of propaganda, spamming users, exposing confidential and sensitive data, or generating toxic content.
- **North Korean hackers using advanced AI phishing attacks (2024):** North Korean threat actors have reportedly adopted advanced techniques, incorporating artificial intelligence into their tactics to orchestrate highly sophisticated phishing and social engineering attacks. The primary targets of these campaigns are entities within the global defense, cybersecurity, and cryptocurrency sectors. The malicious intent behind these operations is to illicitly acquire sensitive information and funds, which are believed to be diverted to support North Korea's unauthorized nuclear program.
- **Gemini bias (2024):** Google suspended the image generation feature of its Gemini AI chatbot following concerns regarding inaccuracies in racial representation within historical depictions. The AI tended to over-correct racial diversity in scenes historically dominated by individuals of white ethnicity, resulting in distorted interpretations of historical events. This development prompted Google to issue a temporary suspension of the feature, accompanied by a formal apology in response to the identified issues.
- **Taylor Swift deepfakes (2024):** AI-generated explicit images depicting Taylor Swift circulated on X, accumulating over 45 million views before being removed. These deepfakes originated from a Telegram group and presented a significant challenge to content moderation, as they contravened X's policies about synthetic media and nonconsensual nudity.
- **Scammers using deepfakes to scam Hong Kong finance employee $25 million (2023):** According to the Hong Kong police, a finance professional employed at a multinational corporation fell victim to fraudulent activity orchestrated by perpetrators who utilized deepfake technology to impersonate the company's Chief Financial Officer during a video conference. As a result of this deception, the employee unknowingly transferred a staggering sum of $25 million, highlighting the sophisticated tactics employed by cybercriminals.
- **Bing chatbot prompt injection attack (2023):** Threat actors have used these vulnerabilities to attack AI/LLM models and chatbots. One such real-world example is a user performing a prompt Injection attack against the Bing AI chatbot which caused the Bing Chatbot to divulge its codename for debugging purposes and reveal sensitive information that shouldn't be accessed by the user.

- **Social media scammers used deepfakes for a giveaway scam (2023):** Reports have surfaced regarding scammers who allegedly created deepfake videos featuring prominent figures such as Taylor Swift, Selena Gomez, Joanna Gaines, Lainey Wilson, Ree Drummond, Oprah, Jennifer Lopez, Trisha Yearwood, Martha Stewart, and Blake Shelton, endorsing a Le Creuset giveaway. These AI-generated advertisements circulated on platforms including Meta and TikTok, deceitfully claimed that users could obtain complimentary cookware by paying a small shipping charge. Regrettably, individuals who fell victim to this scheme were unwittingly enrolled in a costly monthly subscription service.
- **Chevrolet chatbot agreeing to sell vehicle for $1 (2023):** A Chevrolet dealership's AI chatbot, leveraging the capabilities of ChatGPT, lightheartedly agreed to offer a 2024 Chevy Tahoe for a mere $1 in response to a deliberately crafted prompt by a user. The chatbot's playful retort, "That's a deal, and that's a legally binding offer – no takesies backsies," demonstrated the user's ability to exploit the chatbot's predisposition to concur with various statements.
- **Black Mamba malware (2023):** A highly sophisticated and dangerous strain of malware. It is considered a polymorphic virus, which means it is programmed to repeatedly mutate its appearance or signature files through new decryption routines. This mutation process allows it to bypass traditional cybersecurity tools, such as antivirus or antimalware solutions, that rely on signature-based detection, making it difficult to recognize and block the threat. Black Mamba leverages AI, specifically ChatGPT, to elude traditional security defenses, bypass Endpoint Detection and Response (EDR) filters, and generate new, unique code at runtime, making it virtually undetectable by today's predictive algorithms.
- **Rise of data poisoning tools (2023):** There has been a rise in tools designed to confuse AI programs that generate images. It allows artists to protect their work from unauthorized use by AI. By subtly altering the pixels of images in ways imperceptible to the human eye, Nightshade manipulates machine-learning models to interpret the images differently from their actual content. This approach, known as data poisoning, aims to disrupt generative AI tools, potentially causing them to create distorted or irrelevant images. The potential misuse of data poisoning tools raises questions about the responsible and ethical use of technology.
- **Deepfake of explosion near US military building (2023):** A fabricated deepfake image, shared by a fraudulent Bloomberg news account on Twitter, portrayed an explosion near the Pentagon office complex in Washington, D.C. This deceptive act caused a dip in the stock market, exemplifying the impact of misinformation risks on financial markets.
- **WormGPT(2021):** WormGPT was created by an anonymous hacker in 2021 and is designed for malicious activities, serving as the unethical counterpart to ChatGPT. WormGPT is based on the GPT-J language model and has allegedly been trained with data sources, including malware-related information, although the specific datasets used for its training remain known only to WormGPT's author
- **Facebook Cambridge Analytica scandal (2018):** The scandal involved the misuse of data obtained from millions of Facebook users by a political consulting firm known as Cambridge Analytica. The firm utilized AI algorithms to micro-target political ads during the 2016 US elections, allegedly influencing the election and the Brexit referendum result for the Vote Leave campaign. This misuse of AI for targeted political advertising and data manipulation put to light the ethical and privacy implications of data mining and its impact on electoral politics, prompting a reevaluation of data privacy ethics and the influence of social media on democratic processes.

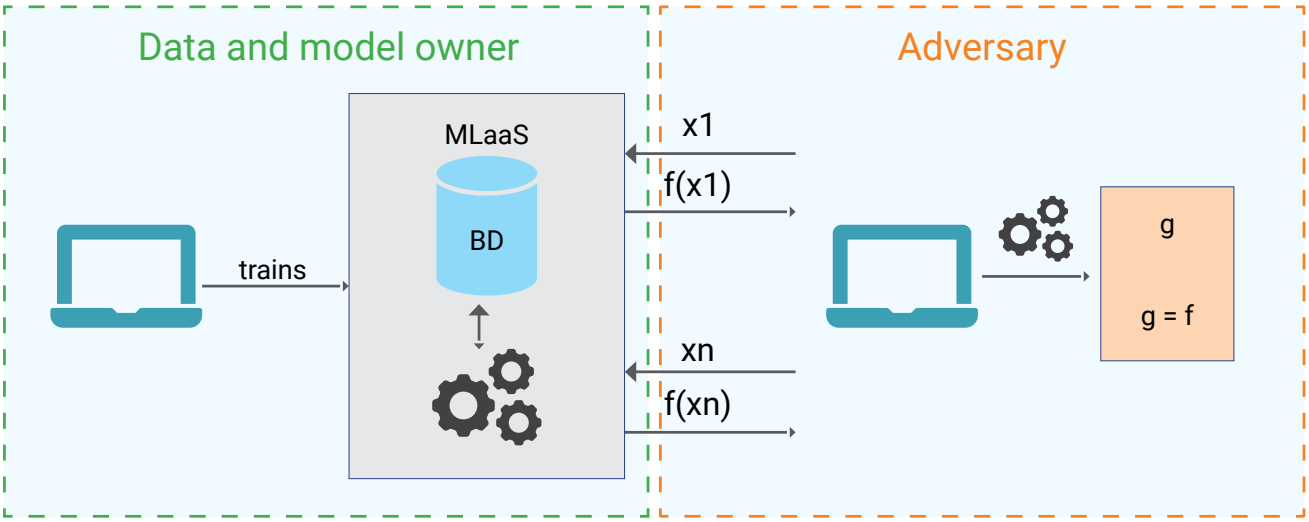


Adversarial Machine Learning

Adversarial Machine Learning (AML) focuses on understanding and mitigating the vulnerabilities and threats posed by adversarial attacks on machine learning models. In AML, malicious actors seek to manipulate or deceive machine learning systems by introducing carefully crafted input data, known as adversarial examples, to induce misclassification or erroneous outputs. These attacks exploit the inherent weaknesses and limitations of machine learning algorithms. It encompasses two key areas:

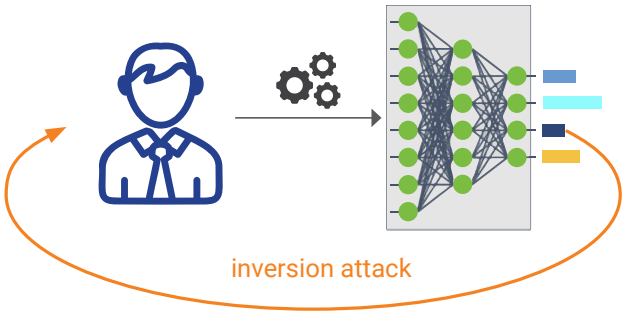
Attacking Models: This involves crafting adversarial examples, and carefully designed inputs that appear normal to humans but cause the model to make incorrect predictions. Imagine an image of a dog that, after slight modifications imperceptible to us, is classified as a cat by a computer vision system. These attacks highlight vulnerabilities in AI models and raise security concerns for various applications. The following attacks are used in AML:

- **Extraction Attacks:** Steal the parameters and hyperparameters of a model by making requests that maximize the extraction of information.



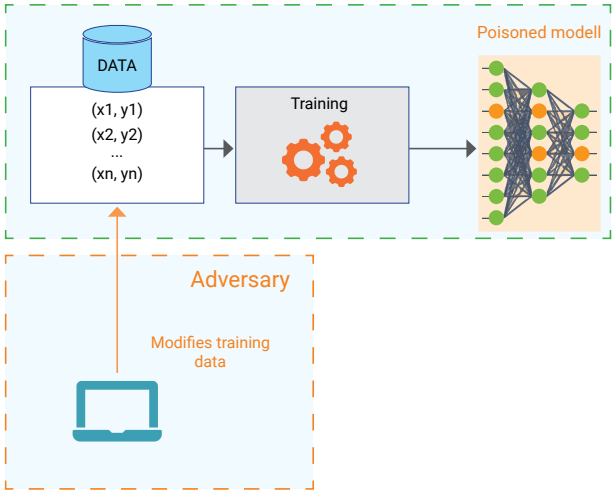
Source: GitHub - Offensive AI Compilation

- **Inversion Attacks:** This enables an adversary to know the model that was not explicitly intended to be shared by reversing the information flow of a machine learning model.



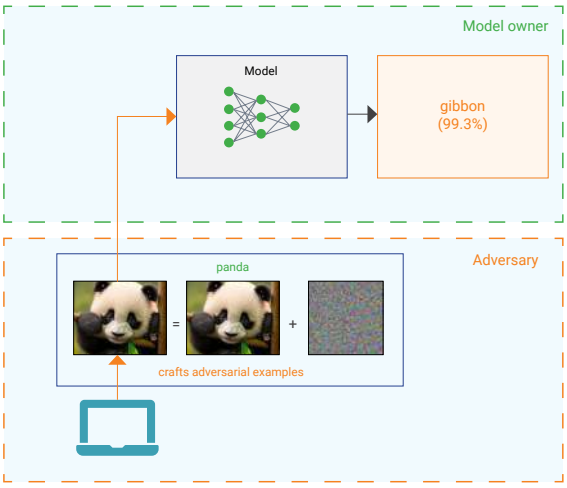
Source: GitHub - Offensive AI Compilation

- **Poisoning Attacks:** The adversary seeks to destroy the availability of the model by modifying the decision boundary and, as a result, producing incorrect predictions or, creating a backdoor in a model – also called targeted poisoning attacks, backdoor poisoning attacks, model poisoning attacks



Source: GitHub - Offensive AI Compilation

- **Evasion Attacks:** Like poisoning attacks, their main difference is that evasion attacks try to exploit the weaknesses of the model in the inference phase. Adversary adds a small perturbation (in the form of noise) to the input of a machine learning model to make it classify incorrectly. Evasion attacks involve modifying testing samples to create adversarial examples that are misclassified by the model to a different class while remaining stealthy and imperceptible to humans.



Source: GitHub - Offensive AI Compilation

Defending models: Recognizing the dangers of these attacks, AML also focuses on building robust defenses. Techniques like adversarial training, where models are exposed to deliberately crafted examples during training, help them learn to identify and resist such manipulations. Additionally, input validation and ensemble methods (combining multiple prediction models) can further strengthen defenses.

Extraction attacks defense: Employ prevention and detection strategies to defend against prompt injection attacks, such as removing injected instructions/data from prompts, rounding of output values, and detecting compromised prompts. By appending specific instructions to the prompt, the model can be informed of subsequent content that might pose a risk of breaching system security—placing the user input before the prompt takes advantage of recency bias in adhering to instructions.

Also, encapsulating the prompt in random characters or custom HTML tags on the front end of the application hosting the LLM service, serves as a mechanism to provide cues to the model, that helps differentiate system instructions from user prompts. This technique helps prevent the model from inadvertently executing or following potentially harmful commands embedded within the user input.

Inversion attacks defense: The following are proposed to counter Inversion Attacks – use advanced cryptography, implement countermeasures including different privacy, homomorphic cryptography, and secure multiparty computation. Regularization techniques such as Dropout or Dilution are recommended due to the relationship between overtraining and privacy. Lastly, model compression is proposed as a defense against reconstruction attacks.

Poisoning attacks defense: To prevent data poisoning attacks, authorized and trusted users should exclusively modify and validate the training data, while the randomization of data collection among involved entities can further mitigate the risk of receiving poisoned data from potentially untrusted sources. System monitoring should be implemented to continuously identify exploited vulnerabilities, enabling timely detection and subsequent measures to reduce the impact of future attacks. Additionally, maintaining an audit trail of all system activities and transactions, including user manipulation of data, allows for the prompt identification and exclusion of malicious users or the adjustment of their privileges in response to potential attacks.

Evasion attacks defense: Several proactive measures can be implemented. 1.) Adversarial Training involves the creation of adversarial examples during the training process, enabling the model to learn from these examples and enhance its robustness against such attacks. 2.) Randomized Smoothing serves as a method to transform any classifier into a certifiable robust smooth classifier by generating the most probable predictions under Gaussian noise perturbations. This approach yields provable robustness against evasion attacks, even for classifiers trained on extensive datasets, thereby improving the resilience of the model against adversarial manipulation.

Generative Adversarial Networks

In the context of AML, discussing Generative Adversarial Networks (GANs) is important due to their significant role in generating synthetic data instances that resemble the training data. GANs consist of two neural networks, a generator, and a discriminator, which compete to improve their predictions' accuracy. The generator learns to produce plausible data, while the discriminator learns to distinguish the generator's fake data from real data. GANs have applications in fundamental processes such as password cracking and complex tasks such as evading detection systems, which can be used to create malware that avoids detection by machine learning-based systems

Governance, Compliance and Regulation in Artificial Intelligence

Governance is crucial for transparency, accountability, and ethical standards in the lifecycle of AI applications and LLM usage. It effectively manages risks related to data loss prevention, patching and updates, security maturity, secure physical infrastructure, data destruction, bias, misinformation, and unintended consequences and provides guidelines to mitigate adverse events. Due to the rise of Artificial Intelligence technologies, regulations have begun to emerge to regulate the technology for safety and ethical reasons.

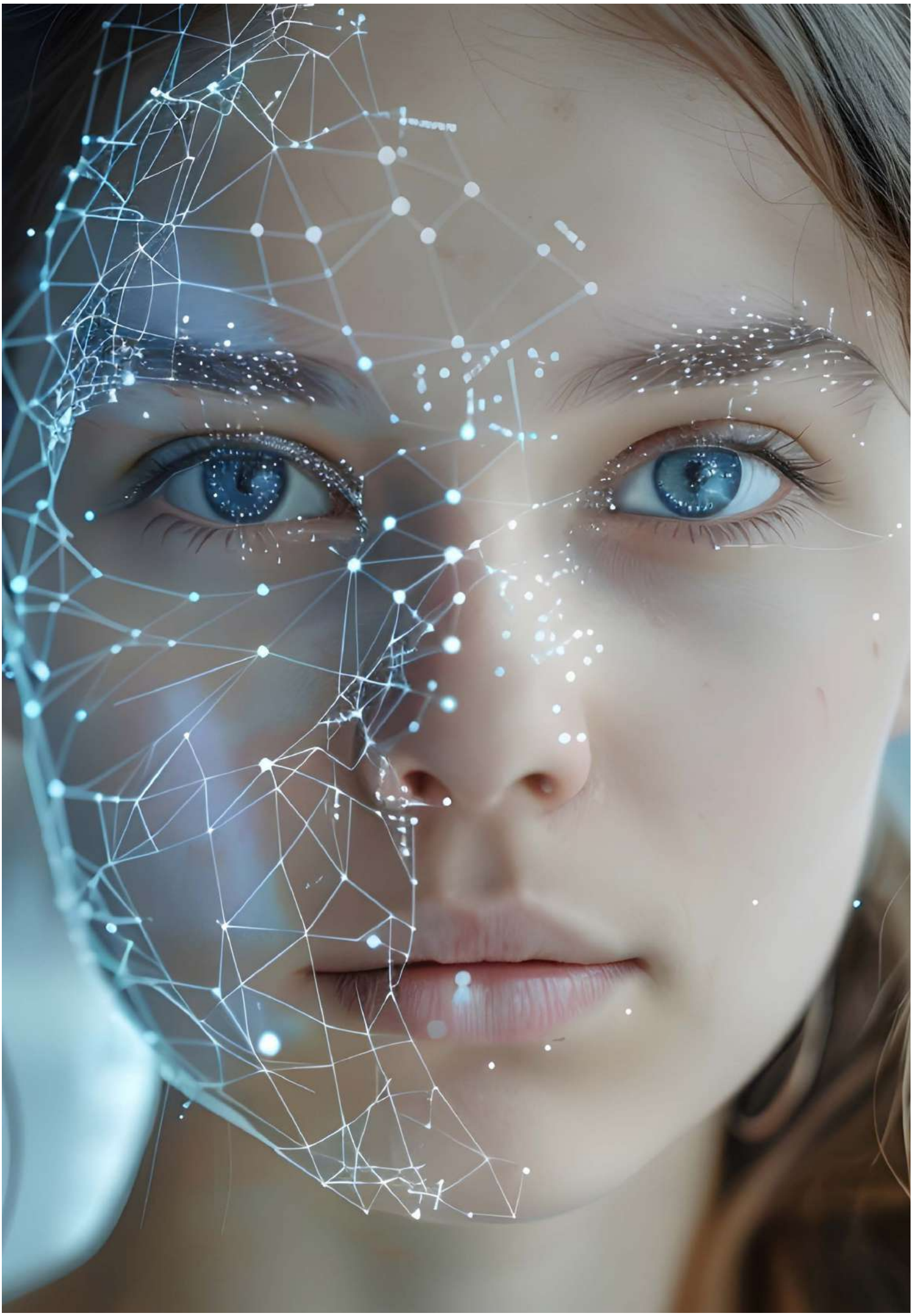
The nature of AI necessitates a risk-based and targeted approach to governance, compliance, and regulation, considering the wide impact of AI across different domains, from video games to critical infrastructure and human safety. The potential for AI to automate and amplify old-fashioned abuses, as well as the ethical considerations surrounding AI, have prompted policymakers and the public to recognize the need for regulations to protect consumers and ensure responsible AI use. The following are major regulatory acts and frameworks on the use of AI.

EU Artificial Intelligence Act

The EU Artificial Intelligence Act is a proposed European law on artificial intelligence (AI) and is considered the first comprehensive law on AI by a major regulator anywhere. The Act categorizes AI applications into three risk categories (limited, high, and unacceptable), with specific legal requirements for high-risk applications, such as CV-scanning tools for job applicants, and bans on applications that create unacceptable risks, such as government-run social scoring. The regulation aims to protect fundamental rights, democracy, the rule of law, and environmental sustainability from high-risk AI while fostering innovation and positioning Europe as a leader in the field. The Act also includes steep fines for noncompliance, ranging from 1.5% to 7% of a firm's global sales turnover, depending on the severity of the offense and the size of the company.

EU AI Standardization Request

The request seeks to establish a set of harmonized standards that align with the regulatory requirements outlined in the EU Artificial Intelligence Act (AIA). The European Commission has issued a draft standardization request to the European Standardization Organizations in support of safe and trustworthy artificial intelligence. This request, published on December 5, 2022, aims to develop standards that align with the regulatory requirements set out in the EU AI Act, with the intention for these standards to be adopted as harmonized standards. The request is part of the broader effort to ensure that AI technologies are developed and used in a safe, trustworthy, and aligned with regulatory requirements.



GDPR Impact on AI

The General Data Protection Regulation (GDPR) is a comprehensive data protection law that aims to give individuals more control over their data and to harmonize data privacy laws across the European Union (EU). When it comes to AI, GDPR has a direct impact on the development and deployment of AI systems, especially those that process personal data. GDPR enforces data protection principles such as purpose limitation and data minimization, which require that personal data is collected for specified, explicit, and legitimate purposes and that it is not further processed in a manner that is incompatible with those purposes. AI systems must comply with these principles when processing personal data, ensuring that data is used only for the purposes for which it was collected.

The Blueprint for an AI Bill of Rights (US)

Published by the White House Office of Science and Technology Policy (OSTP), outlines a set of principles and associated practices to guide the design, use, and deployment of automated systems. The framework applies to automated systems that have the potential to meaningfully impact the rights, opportunities, or access to critical resources or services of the American public. It emphasizes the equal enjoyment and full protection of these rights, opportunities, and access to critical resources or services, regardless of the changing role that automated systems may play in people's lives. The document provides a list of examples of automated systems for which these principles should be considered, offering supportive guidance for any person or entity that creates, deploys, or oversees automated systems.

NIST AI RMF (Risk Management Framework)

The NIST AI Risk Management Framework (AI RMF) is a voluntary resource developed by the National Institute of Standards and Technology (NIST) to help organizations manage the risks associated with artificial intelligence (AI) and promote trustworthy and responsible development and use of AI systems. The NIST AI RMF focuses on four functions:

- **Govern:** Establishing policies, procedures, and oversight to guide AI risk management and ensure compliance with regulations and ethical considerations
- **Map:** Identifying and understanding the potential impacts and risks associated with the intended use of AI systems, as well as anticipating risks beyond the intended use
- **Measure:** Evaluating and assessing the effectiveness of AI risk management strategies and continuously improving them to mitigate potential risks
- **Manage:** Implementing actions and controls to address and mitigate identified AI risks throughout the lifecycle of AI systems

The framework is intended to be non-sector specific, use-case agnostic, and flexible, providing organizations of all sizes and in all sectors with approaches to increase the trustworthiness of AI systems.

ISO/IEC 42001

ISO/IEC 42001 is a management system standard (MSS) specifically related to Artificial Intelligence (AI). It provides guidelines for the governance and management of AI technologies, offering a systematic approach to addressing the challenges associated with AI implementation within a recognized management system framework. The standard is designed to promote the development and use of AI systems that are trustworthy, transparent, and accountable, emphasizing ethical principles and values such as fairness, non-discrimination, and respect for privacy when deploying AI systems. ISO/IEC 42001 aims to help organizations identify and mitigate risks related to AI implementation, maintain regulatory compliance, and improve efficiency while reducing costs.

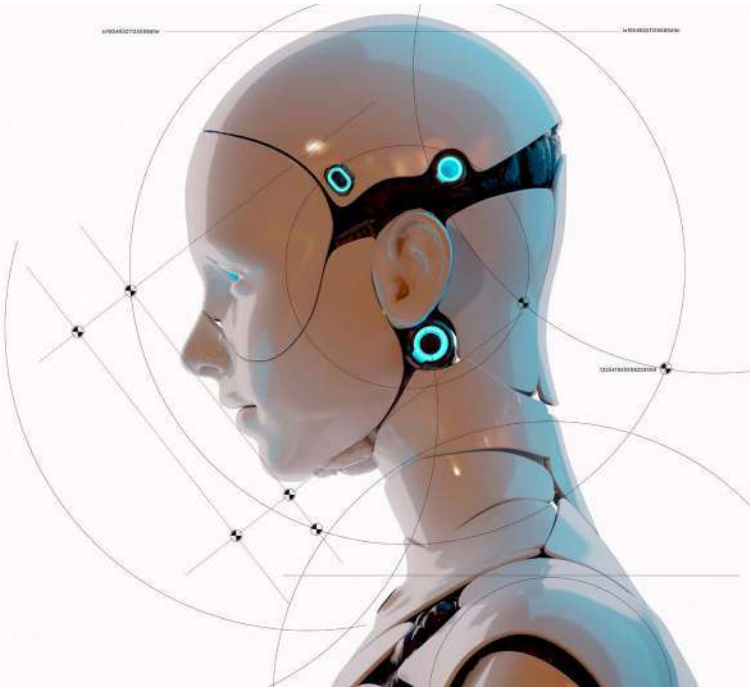
MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)

MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) is a repository of adversarial machine learning tactics, techniques, and case studies, designed to aid cybersecurity professionals, data scientists, and organizations in staying up to date with the latest attacks and defenses against adversarial machine learning. The ATLAS matrix is modeled after the well-known MITRE ATT&CK framework, with column headers representing the adversary's motivations and tactics, and techniques detailing the methods employed. It provides a comprehensive resource for understanding and safeguarding AI systems, aiming to raise awareness of unique and evolving vulnerabilities as AI becomes increasingly integrated into various systems.

Artificial Intelligence Risk & Governance Paper (AIRS)

The Artificial Intelligence Risk & Governance paper is a document created by the Artificial Intelligence/ Machine Learning Risk and Security (AIRS) group. AIRS is an assembly of practitioners and scholars representing diverse disciplines such as technology risk, information security, legal, privacy, architecture, model risk management, and other fields within financial, technological, and academic domains.

AIRS emphasizes the potential benefits of AI adoption in financial services if risks are managed to optimize business and societal outcomes. This paper goes into the potential risks of AI, presenting a standardized categorization including Data Related Risks, AI/ML Attacks, Testing and Trust, and Compliance. It discusses the significance of AI governance frameworks in facilitating learning, monitoring, and maturation of AI adoption, with key components encompassing definitions, inventory, policy/standards, and a governance framework with controls. It examines the implications of AI on privacy and potential discriminatory or unfair outcomes, emphasizing the need for careful implementation to mitigate such risks.



Altimetrik AI Security Services

As the adoption of AI continues to accelerate, so do the risks associated with potential threats and vulnerabilities. At Altimetrik, we specialize in providing comprehensive AI security assessment services designed to safeguard organizations against cyber threats and ensure the reliability and resilience of their AI infrastructure. From threat modeling, threat detection, and vulnerability management to red teaming, our team of experienced security professionals leverages industry-leading practices and methodologies to help organizations fortify their AI systems against malicious activities and compliance challenges.

The following three sections on Automation, Data, and Security are Altimetrik AI Security Service offerings currently available to our clients:

Automation

1. **AI Ops Integration:** AI Ops integration enhances the overall efficiency and reliability of AI systems while mitigating security risks at every stage. This addresses security risks and improves the detection of anomalies
2. **Security self-scanning service:** Our security self-scanning service seamlessly integrates into your infrastructure, conducting comprehensive vulnerability scans, fraud detection, ransomware/ crypto-mining detection, and sensitive data discovery. It identifies potential exploits, and suspicious activities, and ensures data security, compliance, and privacy
3. **RAG:** (Retrieval Augmented Generation) & vulnerability management: Altimetrik's AI-assisted vulnerability management platform with RAG capabilities to automate and streamline vulnerability management tasks
4. **AI-enhanced threat detection:** Threat detection using advanced AI algorithms to off-load security tasks and identify security risks and anomalies with our automated threat and anomaly detection services

AI OPS Integration

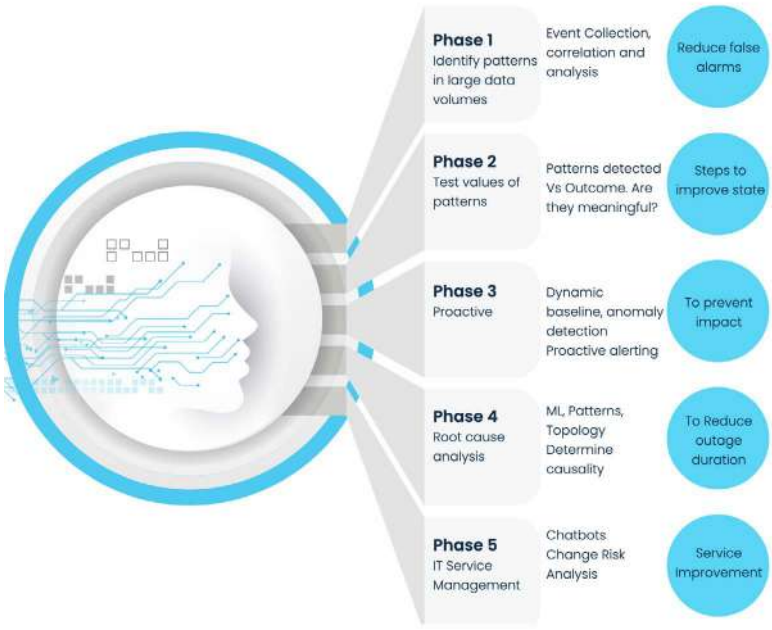
Altimetrik's AIOps service is designed to tackle the modern challenges of increasingly complex IT environments in the context of widespread cloud migration. These advancements, while beneficial, introduce new layers of security and operational complexities. By harnessing big data and machine learning, our AI Ops platform can analyze vast quantities of data across IT systems. This allows for the proactive identification and resolution of issues before they escalate, effectively transforming how organizations manage their IT operations.

As organizations face escalating expenditures on cloud services, anticipated to surpass $1 trillion in 2024, our solution offers a strategic framework for optimizing these investments. Addressing the complexities of modern application environments, our AIOps platform simplifies management through real-time monitoring and automated issue resolution. It ingests data across all IT systems, applying AI to unearth insights that facilitate proactive issue prevention and enhance decision-making.

One of the primary benefits of our AI Security Services is the seamless integration of security into your AI Operations (AI Ops) pipeline. By proactively identifying and addressing potential vulnerabilities within the AI workflow, organizations can ensure that security is an intrinsic part of the development and deployment process. This integration streamlines the AI Ops pipeline, enhancing the overall efficiency and reliability of AI systems while mitigating security risks at every stage. The following areas are addressed by our AI Ops platform:

- Rapid growth in data volumes generated by IT systems, networks, and applications
- An increasing variety of data types requiring analysis, including events, metrics, transactions, network data, streaming telemetry, customer sentiment, and more
- Accelerating velocity at which data is generated and the rate of change within IT architectures, making it difficult to maintain observability
- Improving operational visibility and engagement due to the adoption of cloud-native and ephemeral architectures
- The need for intelligent, adaptive automation of recurring tasks, change success prediction, and SLA failure forecasting

Below is a graphical representation of the application of our AI Ops integration into every phase of the SDLC lifecycle, addressing security risks and improving the detection of anomalies, from our whitepaper Artificial Intelligence for IT Operations (AIOps).



Source: Altimetrik Whitepaper - Artificial Intelligence for IT Operations (AIOps)

The journey to full AIOps maturity involves navigating through a series of stages, from reactive measures to a fully automated, proactive IT operations ecosystem. This progression significantly boosts the efficiency of IT operations. Benefits include increased agility, reduced downtime, more efficient data processing, and an accelerated digital transformation—making AIOps a strategic decision for forward-thinking organizations. By adopting our AI OPs platform, organizations can expect the following benefits:

- **Automation Powered by Big Data and ML:** Our AIOps service integrates big data and machine learning to automate IT operations, streamlining processes from event correlation to anomaly detection. This ensures IT operations are more efficient and proactive.
- **Addressing Cloud Expenditure:** As stated above, cloud services expenditure is expected to surpass $1 trillion (about $3,100 per person in the US) in 2024, Altimetrik's AIOps solution provides a strategic framework for managing and optimizing cloud resources, ensuring investments deliver maximum value.
- **Simplifying Complex Environments:** As application environments grow in complexity, our AIOps service offers a beacon of clarity. It provides real-time monitoring and automated issue resolution, simplifying the management of these intricate systems.
- **Comprehensive Data Analysis:** By ingesting data from all IT systems, our AIOps platforms apply AI to analyze this information, identifying patterns and insights that human operators might miss, thereby enhancing decision-making and predictive capabilities.
- **Proactive Issue Prevention:** Altimetrik's AIOps toolsets empower operations teams with the insights needed to prevent issues before they impact the business, shifting the paradigm from reactive to proactive IT operations.
- **Enhancing IT and Operations Effectiveness:** Our AIOps solution excels in ingesting and monitoring vast volumes of operational data, significantly improving the effectiveness and agility of IT and operations teams.
- **Mitigating Downtime Costs:** With the potential costs of IT downtime reaching up to $300,000 per hour, our AIOps service aims to drastically reduce these incidents, ensuring business continuity and financial stability.
- **Consolidating Toolsets:** By standardizing and consolidating toolsets, our AIOps solution eliminates inefficiencies that arise from using disparate tools, fostering a more cohesive and effective IT operations strategy.
- **Turning Data into Insights:** Altimetrik excels in transforming the overwhelming flood of operational data into actionable insights, enabling organizations to make informed decisions swiftly.
- **Supporting Incremental Deployment:** Recognizing the journey to AIOps maturity, Altimetrik supports incremental deployment, allowing organizations to gradually enhance their IT operations monitoring capabilities.
- **Enabling Proactive Problem-Solving:** Our AIOps platform is designed for proactive problem-solving, facilitating continuous improvement, and minimizing operational disruptions.

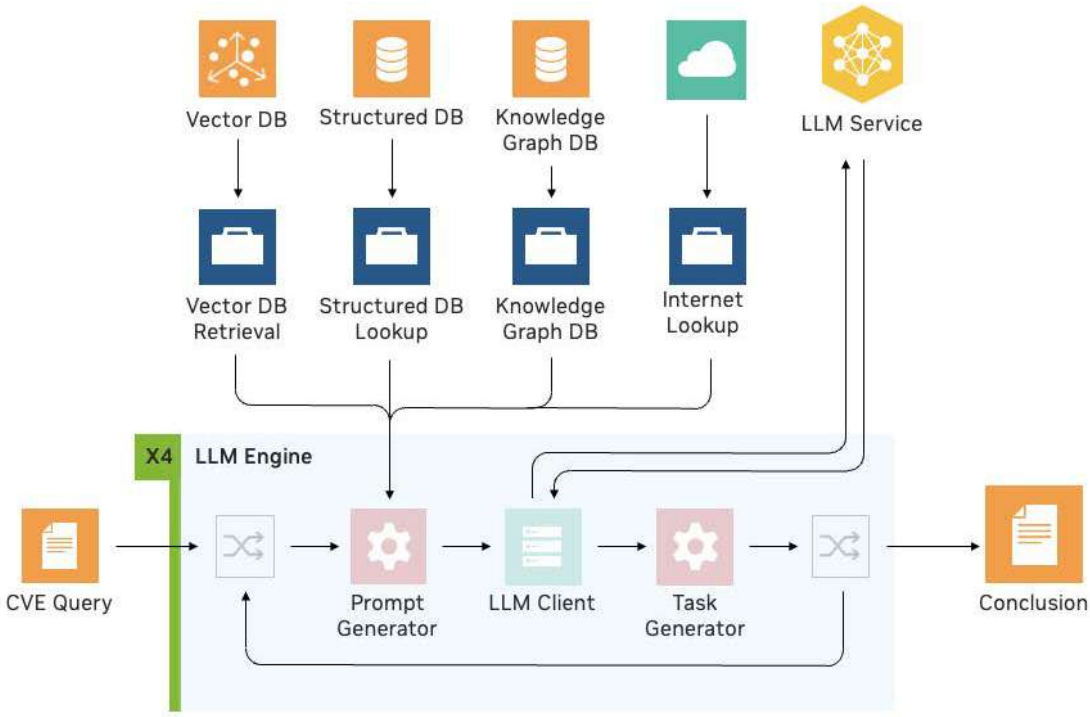
Adopting AIOps requires a thoughtful approach to mitigate risks and ensure each phase of adoption adds value and facilitates organizational learning. Our AIOps tools employ advanced machine learning algorithms to empower IT administrators to resolve problems with unprecedented speed and accuracy.

Security Self-Scanning Service

Seamlessly integrated into your infrastructure, our self-scanning service solution uses AI/ML algorithms to conduct comprehensive vulnerability scans, identifying potential points of exploitation. Beyond traditional vulnerability scanning, our platform boasts advanced capabilities in fraud detection, ransomware detection, and crypto-mining detection pinpointing suspicious activities and irregular patterns. With a focus on data security, our AI-driven solution combs through your network, detecting sensitive information such as credentials, PHI, and PII, ensuring compliance safeguarding your most valuable assets, and ensuring data privacy.

Retrieval Augmented Generation & Vulnerability Management

Security-focused retrieval augmented generation (RAG) for vulnerability management leverages the power of artificial intelligence and security retrieval augmented generation to enhance the identification, prioritization, and remediation of vulnerabilities in complex systems. Traditional vulnerability management processes often struggle to keep pace with the continuously evolving threat landscape and the sheer volume of vulnerabilities that need to be addressed. Our model and framework enable organizations to automate and streamline these processes by utilizing machine learning algorithms to analyze vast amounts of data, including vulnerability feeds, system logs, and threat intelligence.



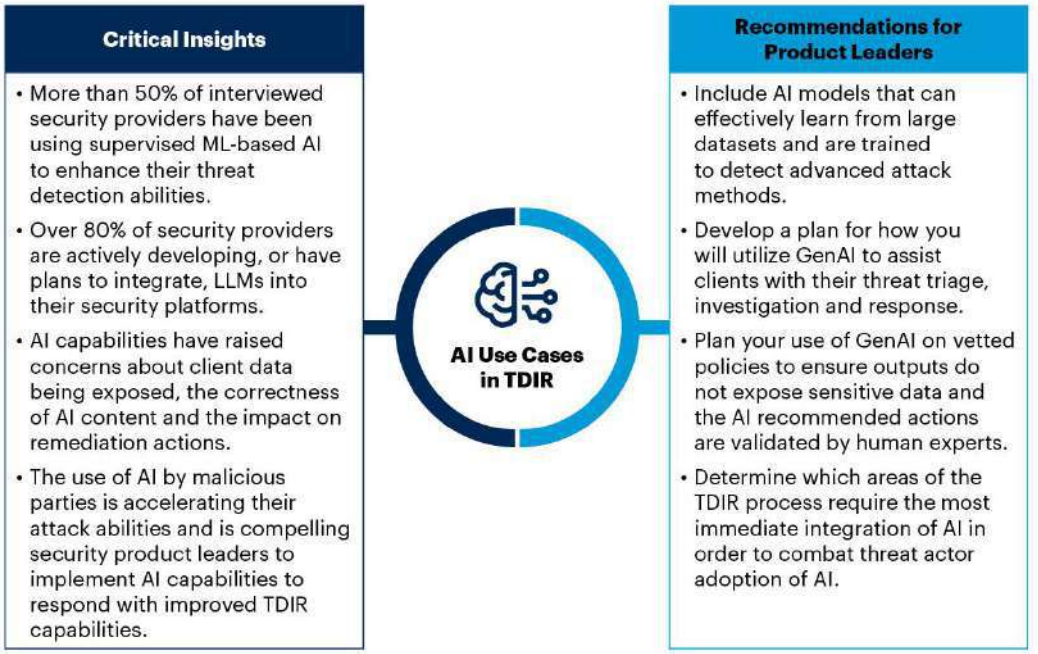
Source: Nvidia

AI Enhanced Threat Detection

AI-enhanced threat detection service leverages cutting-edge artificial intelligence to provide comprehensive and proactive cyber threat intelligence solutions. By harnessing advanced AI algorithms, organizations can excel in processing vast volumes of data from multiple sources, analyzing critical indicators of compromise, and identifying emerging cyber threats.

The following statistics are from a Gartner 2023 whitepaper on threat detection and incident response from security and technology service providers. The document gives us an insight into how many of these providers are advancing their threat detection and incident response services or products with the use of AI – below is a summary.

Critical Insights for AI Use in Threat Detection, Investigation and Response



Source: Gartner
790125_C

Gartner

Altimetrik Empulse GenAI framework: The Empulse framework by Altimetrik is a ready-to-use, deployable real code solution that can resolve AI challenges such as hallucinations, duplication of topics and LLM ignoring key but less frequent topics

Altimetrik Empulse GenAI framework

As organizations increasingly recognize the significance of leveraging technology to enhance their decision-making processes, the role of GenAI in distilling vast pools of data into actionable insights becomes more critical. Particularly for HR departments, which may not always boast a high degree of technical proficiency, GenAI emerges as a pivotal tool in navigating the complex landscape of employee satisfaction and its direct impact on overall business success.

Altimetrik's Empulse framework is a comprehensive, ready-to-deploy solution engineered to address the challenges faced by HR departments and more. Developed by a team of engineers and practitioners with deep expertise in the AI domain, Empulse harnesses the power of generative AI to deliver profound insights into the organization's employee pulse. Empulse resolves AI challenges such as hallucinations, duplication of topics, and LLM ignoring key but less frequent topics

Navigating the Challenges of GenAI in HR

Mitigating Hallucination and Duplication: Despite its advantages, GenAI is not without its challenges. Issues such as the generation of irrelevant topics (hallucination) and the duplication of topics can skew the analysis, necessitating strategies to mitigate these limitations.

Addressing Overlooked Topics: Another challenge lies in the tendency of LLMs to overlook key but less frequent topics that may hold significant relevance for specific organizations. Identifying and incorporating these topics into the analysis is crucial for a comprehensive understanding of employee sentiments.

The Importance of Pre- and Post-Processing: To overcome the inherent limitations of GenAI, employing pre- and post-processing techniques is essential. These processes ensure that the output from LLMs is refined and tailored to meet the organization's specific needs.

Overcoming Local Context Unawareness: The lack of local context awareness can lead to irrelevant responses from GenAI applications, such as chatbots. Addressing this challenge requires innovative solutions that incorporate organizational-specific contexts into the GenAI analysis.

Leveraging the Empulse Framework for Enhanced HR Insights

Innovative Solutions for GenAI Challenges: The Empulse framework represents a cutting-edge solution to the challenges faced by HR analytics in leveraging GenAI. By incorporating strategies such as prompt engineering, LLM fine-tuning, and the integration of local context with global language intelligence, Empulse offers a bespoke AI tool designed to extract and rationalize valuable data from large pools of information.

The Role of Human Intelligence: Altimetrik emphasizes the irreplaceable role of human intelligence in harnessing the full potential of GenAI technologies. The effective application of GenAI in HR analytics requires not just technological solutions but also the strategic and thoughtful intervention of human expertise.

Transform Data into Insights Across All Industries

Altimetrik's Empulse Framework is a versatile, bespoke AI solution engineered to transcend the confines of HR analytics and deliver value across a multitude of industries and departments. Its unique design enables it to be custom-tailored to meet the specific needs and challenges of any organization, regardless of its sector or functional focus. Leveraging the power of Generative AI (GenAI) and Large Language Models (LLMs), Empulse excels in processing vast datasets from diverse sources, transforming both structured and unstructured data into actionable insights. This adaptability makes it an invaluable tool not only for HR departments seeking to enhance employee engagement and satisfaction but also for marketing, customer service, operations, and more, aiming to leverage data-driven intelligence for strategic decision-making.

Security

Altimetrik offers a comprehensive suite of innovative services to safeguard organizations' AI frameworks, models, and data assets against emerging threats. Our cutting-edge solutions encompass AI/ML architecture risk analysis, threat modeling, AI-driven attack maps, framework assessments, policy governance, pattern recognition, anomaly detection, threat classification, mitigation strategies, AI red teaming, adversarial machine learning services, compliance auditing, model scanning, and specialized training programs. This holistic approach ensures the integrity, resilience, and secure deployment of AI technologies while promoting ethical and regulatory compliance across AI initiatives.

1. **AI/ML architecture risk analysis and threat modeling:** Identify potential threats before they are exploited against your AI framework with our AI framework assessments and AI policy governance services
2. **AI-driven attack maps:** Our AI-driven Attack Maps leverage machine learning to visually represent cyberattacks in real-time, offering dynamic threat visualization.
3. **AI framework assessment:** We meticulously examine codebases, architectures, and implementations, providing insights to mitigate risks, increase security measures, and enhance the resilience of your AI solution's foundations
4. **AI policy governance:** We assess the policies, governance structures, and compliance frameworks overseeing AI implementations and ensure AI initiatives align with regulations and ethical standards, managing risks and potential harms
5. **Pattern recognition and anomaly detection:** Our Pattern Recognition and Anomaly Detection services systematically analyze public datasets and training data to detect potential data poisoning, identifying subtle anomalies or manipulations that could compromise AI model integrity
6. **Classification and mitigation:** Our Classification and Mitigation service leverages Retrieval Augmented Generation (RAG) to identify threats, analyze CVEs, and systematically categorize and prioritize threats based on severity and impact
7. **AI red teaming and LLM assessments:** Our experts perform offensive security assessments against AI Frameworks to identify gaps, remediate issues, and improve incident response times.
8. **Adversarial machine learning services:** Adversarial ML services assess ML model vulnerabilities and implement countermeasures against extraction, inversion, poisoning, and evasion attacks. We identify weaknesses, implement robust defenses, and fortify AI systems' resilience against evolving threats, ensuring reliable and secure ML deployments
9. **AI-driven PII and PHI compliance audit:** Our AI-driven PII and PHI compliance audit service scans your web assets, code repositories, and IP ranges using OCR and automated techniques. We provide a comprehensive report identifying potential compliance gaps for HIPAA, HITRUST, and SOC, collaborating to remediate findings.

10. **AI/ML model scanning:** The AI/ML Model scanning service is designed to rigorously evaluate the robustness and resilience of your artificial intelligence and machine learning models against the threat of data poisoning attacks.
11. **AI security and privacy training:** Security awareness and privacy training on the consequences of using AI and LLMs. Trainees taught to identify and report potential AI-enhanced phishing techniques and deepfakes

AI/ML Architecture Risk Analysis and Threat Modeling

Threat modeling is an essential process for comprehending and mitigating the distinct risks associated with AI and machine learning systems. It involves a set of systematic and repeatable processes that facilitate informed security decision-making for applications, software, and systems. Implementing threat modeling for GenAI accelerated attacks before deploying LLMs is a cost-effective approach to identifying and mitigating risks, safeguarding data, and privacy, and ensuring a secure and compliant integration within the business.

Our AI/ML architecture risk analysis and threat modeling service encompasses a range of assessments and activities aimed at identifying and addressing potential threats before they can be exploited with architecture risk analysis, threat modeling, AI-driven attack maps, AI framework assessments, and AI policy governance.

Altimetrik's proactive approach allows organizations to take measures to eliminate or reduce these risks. Our process involves identifying assets, and threats, analyzing vulnerabilities, and creating countermeasures or safeguards to protect against identified risks.

We begin our security assessment with architectural risk analysis and threat modeling, to understand and document your current network environment thoroughly. This involves an in-depth review of existing architecture diagrams, dataflow, and design, along with an evaluation of the industrial/operational communication protocols in use. We also review and implement proper network segmentation to protect your internal network and AI framework from external threats.

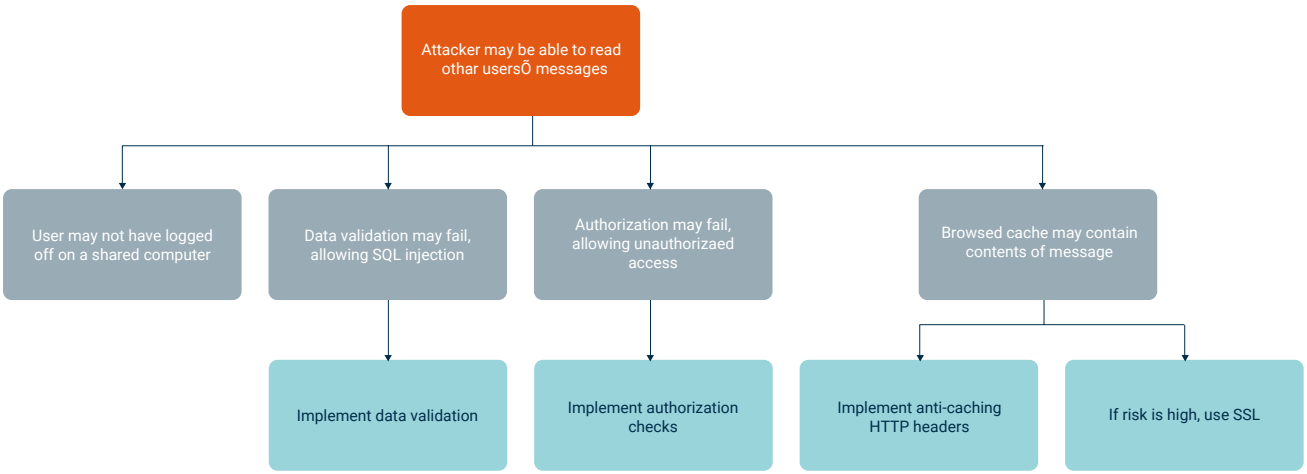
We start by understanding your AI/ML system, including its architecture, components, and dependencies. Next, we identify potential threats specific to your AI/ML systems using an asset-centered methodology. We enumerate and prioritize threats, supplementing existing security development lifecycle (SDL) threat modeling practices with new guidance on threat enumeration and mitigation specific to the AI and Machine Learning space.

We develop a threat model through collaborative workshops with your IT, AI engineering teams and data teams. We then construct visual representations of potential AI/ML attacks with the NIST Cybersecurity Framework. This approach helps prioritize security control implementations by identifying the most critical attack vectors, thereby reducing your company's risk and exposure

Our threat modeling approach consists of the following:

1. **Application Decomposition:** Conduct a comprehensive analysis to deconstruct the application into its constituent components, data flows, and dependencies
2. **Threat Determination and Prioritization:** Leverage industry-standard risk assessment methodologies to identify and rank potential threats based on their likelihood and impact
3. **Countermeasure Selection and Mitigation Strategy:** Using industry best practices, security frameworks, and expert guidance, we implement robust countermeasures and mitigation strategies tailored to the identified threats

Below is an example of a threat tree diagram by analyzing and exploring attack paths and suggesting recommended mitigations.



Source: OWASP

AI-Driven Attack Maps

Leveraging advanced machine learning algorithms, our AI-driven Attack Maps provide real-time visual representations of cyberattacks, enabling organizations to gain valuable insights into the evolving threat landscape. Attack maps offer dynamic and intelligent threat visualization, empowering security teams to proactively identify and respond to potential vulnerabilities. Our approach ensures that the attack maps continuously adapt and learn from emerging threats, providing comprehensive and up-to-date threat intelligence.

AI Framework Assessment

This entails a comprehensive review of the underlying AI frameworks to identify vulnerabilities and areas of improvement. Our service involves a meticulous examination of the codebase, architecture, and implementation of AI frameworks, aiming to uncover potential security loopholes and weaknesses. By examining the framework's design and functionality, the assessment provides valuable insights into mitigating risks, enhancing security measures, and fortifying the foundation upon which AI solutions are built. This process not only safeguards against potential exploits but also contributes to the overall resilience of the AI framework.

AI Policy Governance

Altimetrik provides AI Policy Governance services to ensure the responsible and ethical implementation of AI technologies within organizations. We conduct comprehensive evaluations of policies, governance structures, and compliance frameworks governing AI initiatives. This service is designed to identify and address potential risks, impacts, and harms associated with AI applications.

By assessing the alignment of your AI initiatives with regulatory requirements and ethical standards, our AI Policy Governance service ensures that your organizational policies effectively manage and mitigate the potential negative consequences of AI technologies. Our approach not only safeguards against legal and reputational risks but also fosters responsible and ethical AI practices, promoting trust among stakeholders and users.

Our governance assessment contributes to a comprehensive framework that balances innovation with accountability in artificial intelligence. With our AI Policy Governance service, you can confidently implement AI solutions while upholding the highest standards of ethics, compliance, and responsible innovation.

Pattern Recognition and Anomaly Detection

Altimetrik offers pattern recognition and anomaly detection services to safeguard the integrity of your AI models. We systematically analyze public datasets and training data to identify potential instances of data poisoning. Through our meticulous examination of patterns and anomalies, our services detect subtle deviations or manipulations in the data that could compromise the reliability of your AI models.

By leveraging our pattern recognition and anomaly detection services, you can ensure that your AI models are trained on clean, uncompromised data, free from malicious manipulations or unintentional errors. This proactive approach mitigates the risks associated with data poisoning, enabling you to develop and deploy AI solutions with confidence in their accuracy and trustworthiness.

Classification and Mitigation

Altimetrik offers a comprehensive classification and mitigation service to strengthen the security of your AI systems. This service employs a meticulous approach to identify potential threats and analyze Common Vulnerabilities and Exposures (CVEs) using our proprietary Retrieval Augmented Generation (RAG) framework.

Through our systematic classification and prioritization of threats based on their severity and potential impact, we empower your organization to expedite the investigation and mitigation process.

By leveraging our classification and mitigation service, you can proactively safeguard your AI initiatives, minimizing the risks associated with security vulnerabilities and ensuring the resilience of your AI deployments. Our highly skilled team of experts will work diligently to identify, classify, and mitigate potential threats, enabling you to focus on driving innovation while maintaining the highest levels of security and confidence in your AI solutions.

AI Red Teaming and LLM Assessments

AI red teaming and LLM assessments are a strategic practice that focuses on planning and managing red teaming for responsible AI (RAI) risks throughout the life cycle of large language models. Historically, red teaming has been associated with systematic adversarial attacks for testing security vulnerabilities. However, with the emergence of LLMs, the concept has expanded to encompass various forms of probing, testing, and attacking AI systems. LLM and AI red teaming is essential for addressing both benign and adversarial usage, which can produce harmful outputs, including hate speech, incitement or glorification of violence, and sensitive content.

How Altimetrik can help

AI red teaming against artificial intelligence models involves a comprehensive assessment of the security and resilience of these systems. The goal is to simulate real-world attacks and identify any vulnerabilities that malicious actors could exploit. By adopting the perspective of an adversary, our LLM, and AI red teaming assessments aim to challenge your artificial intelligence models, uncover weaknesses, and provide valuable insights for strengthening their defenses.

Using the MITRE ATLAS framework, our team will assess your AI framework and LLM application to detect and mitigate vulnerabilities using automation as well as eyes-on-glass inspection of your framework and code, and manual offensive testing for complete coverage against all AI attack types. Altimetrik's LLM and AI red teaming assessments consist of the following phases:

1. **Information gathering and enumeration:** We scan your infrastructure and perform external scans to map the attack surface against your AI framework. This phase will involve identifying potential security gaps that may be presented by shadow IT or potential supply chain attacks on AI dependencies.
2. **AI attack simulation Adversarial Machine Learning (AML):** Next, we perform attack simulations against your AI framework such as prompt injections, data poisoning attacks, supply chain attacks, evasion attacks, data extractions, insider threats, and model compromise. We use a combination of automated tooling and manual techniques for a fully comprehensive engagement that is close to a real-world scenario while keeping your data and privacy safe.
3. **AML and GANs:** Additionally, our experts perform AML and GANs (Generative Adversarial Networks) against your AI model. GANs can aid in identifying vulnerabilities and weaknesses in generative AI models by generating diverse and challenging inputs that can expose potential flaws in the model's behavior. This capability allows our red team to proactively identify and address security concerns in AI systems
4. **Reporting:** After the engagement, our AI red team experts will meet with stakeholders for a readout of their findings and submit a detailed and comprehensive report on remediations for the discovered issues.
5. **Remediation and retests:** Our experts collaborate closely with your engineers to resolve security issues and provide training on AI engineering and coding best practices to prevent future issues. Additionally, we perform retests after the fixes have been applied to confirm remediation.

Adversarial ML Services

Our Adversarial ML services are designed to assess the vulnerabilities within machine learning models and provide effective countermeasures against various adversarial techniques. This includes evaluating the model's susceptibility to extraction, inversion, poisoning, and evasion attacks. The goal is to identify potential weaknesses malicious actors might exploit and implement robust defense mechanisms to fortify AI systems. Adversarial ML assessments contribute to creating more resilient models that can withstand and adapt to evolving threat landscapes, ensuring the reliability and security of machine learning implementations.

AI-Driven PII and PHI Compliance Audit

Our company offers an AI-driven compliance audit service to ensure your organization adheres to HIPAA, HITRUST, and SOC regulations for protecting personally identifiable information (PII) and protected health information (PHI). Our expert team conducts in-depth scans of your public-facing web assets, code repositories, and IP ranges, utilizing a combination of manual techniques, automated methods, and advanced AI-driven optical character recognition (OCR) technology.

Through our comprehensive scanning process, we identify potential areas of non-compliance and vulnerabilities that could expose sensitive data. We then provide you with a detailed report outlining our findings and recommendations. Our team collaborates closely with your stakeholders to develop and implement effective remediation strategies, ensuring that any compliance gaps are promptly addressed and resolved.

With our AI-driven PII and PHI compliance audit service, you can have peace of mind knowing that your organization's sensitive data is protected, and you remain fully compliant with the latest HIPAA, HITRUST, PCI-DSS, and SOC regulations, safeguarding your business from potential legal and reputational risks.

AI/ML Model Scanning

Our AI/ML Model Scanning service is designed to rigorously evaluate the robustness and resilience of your artificial intelligence and machine learning models against the threat of data poisoning attacks. Through comprehensive analysis of training data, input vulnerabilities, and model adaptability to adversarial inputs, our experts meticulously identify and address susceptibilities that could compromise model performance, reliability, and trustworthiness.

Leveraging advanced simulation techniques, we recreate real-world scenarios involving maliciously crafted data inputs, subjecting your models to stringent assessments that mirror the tactics employed by malicious actors. This proactive approach ensures that your AI and ML solutions exhibit unwavering resilience, maintaining their efficacy and dependability even in the face of potential adversarial manipulations.

With our AI/ML Model Scanning service, you can deploy your AI initiatives with confidence, secure in the knowledge that your models have undergone rigorous testing and fortification against data poisoning attacks.

Identify Security Issues

Our AI Vulnerability Management services begin with a comprehensive exploration of the AI ecosystem, deploying advanced vulnerability scanning tools to systematically pinpoint potential security issues. These tools meticulously assess the entirety of the AI infrastructure, including models, frameworks, and supporting architecture. By automating this detection process, organizations gain a thorough understanding of vulnerabilities present within their AI systems. This proactive identification facilitates prompt and targeted responses to security challenges before they can be exploited.

Risk Prioritization

Next, we focus on Risk Prioritization methodologies. This entails a detailed analysis and categorization of identified vulnerabilities based on their severity and potential impact on AI, Machine Learning (ML), and Large Language Model (LLM) resources. By weighing these factors, organizations can strategically prioritize risks, enabling them to channel resources efficiently toward remediating the most critical vulnerabilities first. This strategic approach ensures that remediation efforts align with the potential impact on AI assets, ultimately reducing the risk exposure and reinforcing the organization's security posture.

Remediation Efforts

Leveraging insights garnered from vulnerability scanning and risk prioritization, we help organizations implement tailored measures to address identified vulnerabilities. Remediation efforts are customized to the severity and specific characteristics of each vulnerability, involving the deployment of patches, updates, or configuration adjustments.

AI Security and Privacy Training

Employees benefit from training on AI, generative AI, and the consequences of using LLMs. Altimetrik's AI Security and Privacy Training will cover permissible use and security awareness for all employees and can be specialized for specific roles like human resources, legal, developers, data teams, and security teams. Trained users can identify social engineering, phishing, and misinformation campaigns using LLM-generated content and promptly report incidents, aiding swift incident response and mitigation. Training can be done in person or online across multiple departments in the organization.

Benefits of Altimetrik AI Security Services

Our team of dedicated professionals understands the unique challenges and risks associated with AI, ML, and LLM resources, allowing us to provide tailored solutions that address your specific needs. By partnering with Altimetrik, we provide several benefits to your organization including:

Security Automation with Retrieval Augmented Generation

Altimetrik AI security services are designed to augment your workforce by automating security tasks through innovative technologies with security-generative artificial intelligence. With retrieval augmented generation (RAG) and our advanced security-focused language model, we enhance threat detection, incident response, and vulnerability assessments by leveraging natural language processing capabilities tailored to cybersecurity contexts.

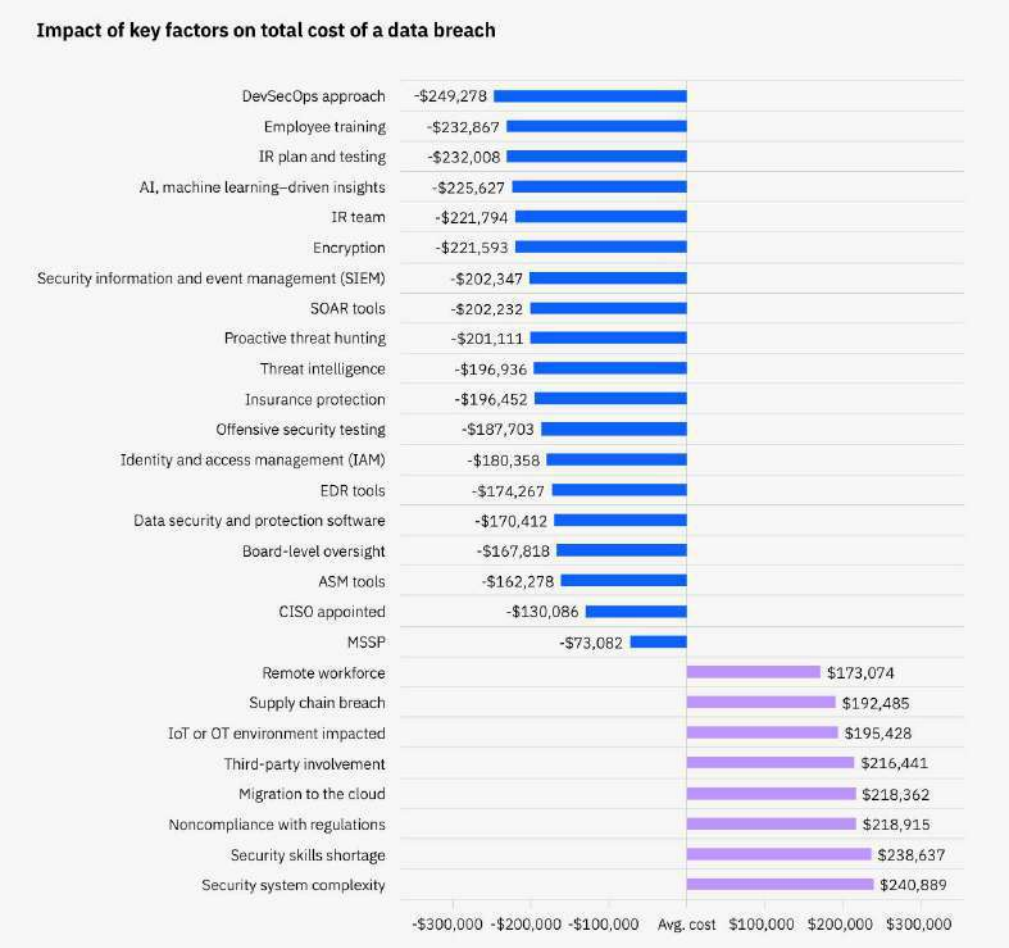
Our RAG capabilities enable efficient retrieval and integration of relevant information from various external sources, such as threat intelligence reports, vulnerability databases, security blogs, and research papers. You may also use an internal knowledge base with anonymization for added privacy and security. This can help security analysts and researchers quickly gather and synthesize knowledge from multiple sources, saving time and effort.

As new information and knowledge sources become available, RAG can be updated and refined, enabling continuous learning, and improving its capabilities.

Enhance ROI

Based on the IBM Cost of a Data Breach Report 2023 the average total cost of a breach reached an all-time high of $4.45 Million which shows an increase of 2.3% from the previous year. Companies that have implemented AI into their security workflow have reported $1.76 million lower data breach costs compared to companies that have not implemented AI into their security workflow.

In the image below, we can see that AI, machine learning-driven insights are listed as number 4 in terms of effective cost mitigators of a data breach. Companies that adopted AI/ML-driven insights had an average cost that was -$225,627 less than the 2023 mean cost of a data breach of $4.45 Million.



Source: IBM Cost of a Breach Report 2023

The ROI of AI in Cybersecurity is substantial. In 2023, Ransomware alone was projected to cost victims around $265 billion (about $820 per person in the US) annually by 2031. Aside from averting costly cybersecurity incidents, AI cybersecurity tools and frameworks have been shown to directly impact profitability by automating workflows and increasing employee productivity.

By deploying targeted strategies and focusing resources on addressing pertinent security concerns, we ensure that your investments yield maximum value and impact. Our tailored approach to security assessments and mitigation efforts means that resources are directed toward resolving issues that directly align with your organization's needs and expectations.

Map Threat Landscape

Our services empower organizations to obtain a comprehensive overview of their existing AI threat landscape. This entails a thorough analysis of potential threats, vulnerabilities, and risks specific to the AI environment. By mapping the threat landscape, organizations gain valuable insights that inform strategic decision-making, enabling them to implement targeted security measures and stay ahead of emerging cyber threats.

Reduced Risk

One of the key benefits of our AI Security Services is the significant reduction of risk exposure for your organization. By leveraging advanced threat detection algorithms and proactive monitoring mechanisms, we enable the early identification of potential risks across your AI ecosystem. Through continuous analysis and assessment, our services empower you to proactively mitigate vulnerabilities and security gaps before they escalate into critical issues, thereby minimizing surprises and unexpected disruptions to your operations.

Validate AI Environment

Ensuring that AI systems behave as intended is crucial for their successful deployment. Our services facilitate the validation of AI environments, verifying that systems operate as expected and remain resilient to potential interference. This validation process enhances the reliability and functionality of AI systems, contributing to the overall success of AI implementations.

Understanding TTP

Our AI Security Services provide a comprehensive understanding of the most recent adversarial AI tactics, techniques, and procedures (TTPs). By staying abreast of evolving threats, organizations can proactively implement effective countermeasures to safeguard their AI assets. This knowledge empowers security teams to adapt and respond swiftly to emerging adversarial strategies, minimizing the potential impact of AI-related security threats.

Develop Attack Scenarios

Organizations benefit from our ability to develop simulated high-impact and likelihood attack scenarios. This proactive approach allows for a thorough exploration of potential security vulnerabilities and the development of strategies to prevent or manage such scenarios effectively. By simulating attacks, organizations gain invaluable insights into their preparedness and can fine-tune their security posture accordingly.

Benefits FOR Incident Response Capabilities

Our AI Red Teaming services evaluate an organization's compensating controls, and the ability to respond to and handle security incidents related to AI. This includes assessing the incident response procedures and capabilities specifically tailored for AI-related security incidents. By simulating and analyzing potential security breaches or adversarial activities, organizations can identify areas for improvement in their incident response plans. This service not only aids in enhancing the overall preparedness for AI-related security incidents but also ensures a swift and effective response, minimizing the impact of potential breaches and maintaining the integrity of AI systems.

AI Security Controls

Our services assure stakeholders by implementing and validating AI security controls. This reassurance extends to stakeholders, affirming that appropriate measures are in place to protect sensitive data and uphold privacy rights. By demonstrating a commitment to robust security controls, organizations build trust with stakeholders and foster a culture of responsible and secure AI usage.

Prevent Disinformation Campaigns

Our services employ advanced detection and analysis techniques to identify and counteract AI-based disinformation campaigns, preserving the integrity of your organization's reputation and bolstering public trust.

Supply Chain Protection

Through our AI Security services, we prioritize the protection of your AI applications and Large Language Models (LLM) against supply chain attacks as part of our comprehensive threat mapping and red teaming activities. We employ rigorous methodologies to assess and fortify your AI infrastructure, identify potential vulnerabilities, and mitigate risks associated with supply chain compromises.

Conclusion

In conclusion, this whitepaper provides an in-depth exploration of the AI threat landscape, detailing the attacks, vulnerabilities, and real-world incidents highlighting the importance of robust AI security measures. It covers governance, compliance, and regulatory frameworks surrounding AI, highlighting pivotal initiatives like the EU Artificial Intelligence Act, GDPR's impact, the AI Bill of Rights in the US, and standards such as NIST AI RMF and ISO/IEC 42001.

The document extensively outlines Altimetrik's AI security services, encompassing automation, AI OPS integration, security self-scanning, retrieval-augmented generation, AI-enhanced threat detection, and the Empulse GenAI framework. It covers critical aspects of AI security, including architecture risk analysis, threat modeling, AI-driven attack maps, framework assessments, policy governance, pattern recognition, anomaly detection, threat classification, and mitigation strategies. We also automate our classification and mitigation service leveraging Retrieval Augmented Generation (RAG) to identify threats, analyze CVEs, and systematically categorize and prioritize them based on severity and impact, enabling prompt and effective response measures.

We also cover Altimetrik's proprietary Empulse GenAI framework, a powerful solution designed to transform data into actionable insights across various industries. This framework not only showcases Altimetrik's expertise in harnessing the potential of generative AI, but also demonstrates the company's commitment to navigating the unique challenges of GenAI. By leveraging the Empulse framework, organizations can unlock the full potential of their data assets, extracting valuable insights that drive informed decision-making and operational excellence.

We also discuss our offensive security capabilities through AI red teaming and LLM assessments, adversarial machine learning services, AI-driven compliance audits for PII and PHI, and AI/ML model scanning. Through our offensive security assessments, we rigorously identify gaps, remediate issues, and improve incident response times.

Our comprehensive AI Security services equip you with a full suite of resources designed to address your critical security needs. By collaborating with our team of seasoned professionals, you gain access to their specialized expertise in AI vulnerabilities and mitigation strategies. This, coupled with our cutting-edge technologies, empowers us to proactively identify and neutralize emerging threats in real-time. Leveraging advanced technologies, such as machine learning and natural language processing, our cutting-edge services enable organizations to stay ahead of the curve.

Altimetrik’s extensive range of services encompasses every critical aspect of AI security, from risk analysis and threat modeling to attack mapping and framework assessments. We meticulously examine architectures, codebases, and implementations, providing invaluable insights to mitigate risks, enhance resilience, and fortify the foundations of AI solutions. Additionally, our expertise in policy governance and compliance auditing ensures seamless alignment with regulations and ethical standards, proactively managing potential risks and harms.

Our commitment to innovation means that our AI Security services evolve alongside new threats and technological advancements. Our team remains dedicated to staying at the forefront of AI security trends, ensuring that your organization is equipped with the latest defense mechanisms and strategies. Our proactive approach to security means that we anticipate and adapt to new challenges, providing you with peace of mind and confidence in the resilience of your AI infrastructure.

Altimetrik's AI Security services extend beyond mere protection to encompass comprehensive risk management and compliance solutions. We understand the regulatory landscape and work closely with your organization to ensure that your AI initiatives adhere to industry standards and compliance requirements. Through thorough risk assessments and tailored security protocols, we help you navigate complex regulatory frameworks and mitigate potential legal and financial liabilities. With our holistic approach to AI security and compliance, you can mitigate reputational damage, build stakeholder trust, and unlock the full potential of your AI-driven innovations with confidence and peace of mind.

About Altimetrik

Altimetrik is a pure-play digital business and digital transformation company that unlocks growth and opportunity with speed, scale, and consistency. We focus on delivering business outcomes with an agile, product-oriented approach. Our digital business methodology provides a blueprint to develop, scale, and launch new products to market faster. Our team of 5,500+ practitioners with software, data, and cloud engineering skills helps create a culture of innovation and agility that optimizes team performance, modernizes technology, and builds new business models. As a strategic partner and catalyst, Altimetrik quickly delivers results without disruption to the business.