Department of the Air Force

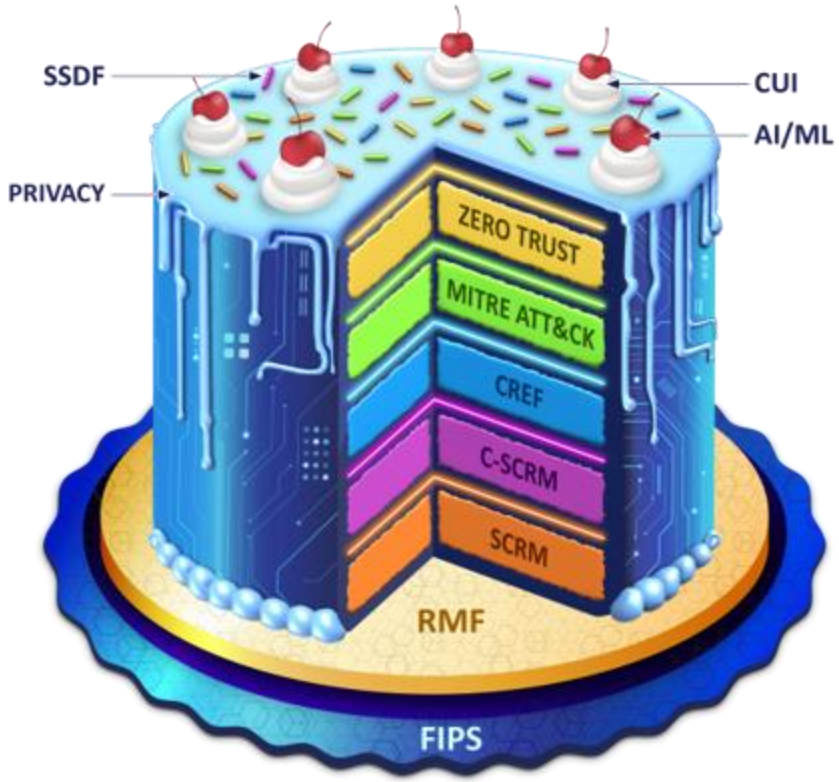
One Team, One Fight!

Artificial Intelligence Large Language Models and Supply Chain Risk

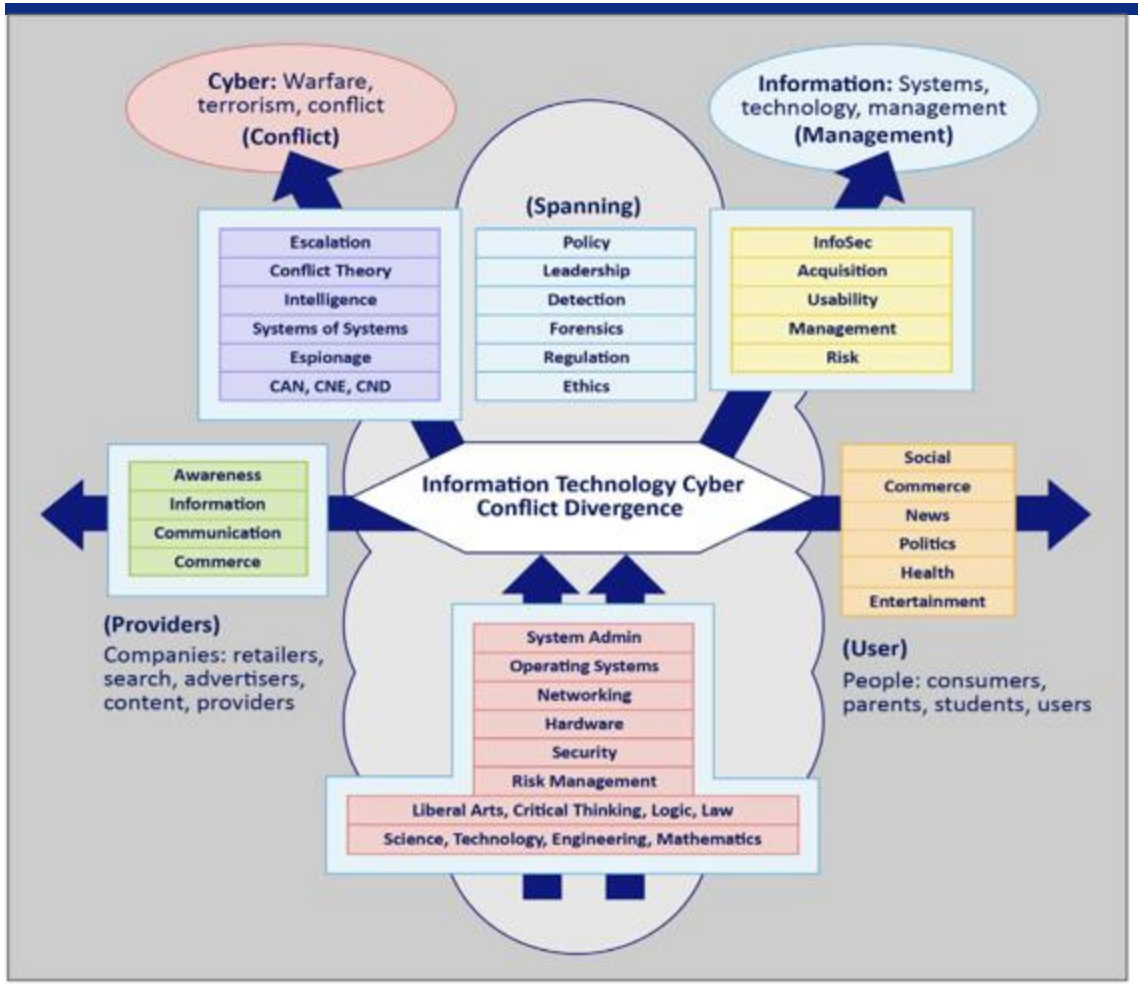
August 2025

What is the Cyber C.A.K.E.

- The Cyber CAKE concept provides a structured, transparent framework that demystifies cybersecurity
- Easy to understand, relatable, useful analogies convey core cybersecurity concepts
- Each layer represents foundation elements of cybersecurity and offers tangible practitioner strategies for grasping the “why” and confidently implement and assess cyber strategies
- Aligned to the Cybersecurity Awareness and Training category within the **Protect** function of the NIST Cybersecurity Framework 2.0 with relevant core “Cookbooks” cited and referenced throughout
- The result of the Cyber CAKE concept is “baked in cybersecurity” and compliance used as strategy to measure resilience



Why the Cyber C.A.K.E.



- Cyberspace is a complex, multidimensional domain that encompasses not only conflict, but also communication, commerce, and societal development
- The knowledge base is inherently fragmented
- There is no true cyberspace “expert”
- Need: transform a daunting obstacle course into a series of manageable steps

Term: Artificial Intelligence

1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.

2) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.

3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.

4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.

5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting (Public Law 115-232, Sec. 238)

Large Language Model (LLM)

- A type of AI model, specifically a deep learning algorithm, trained on vast amounts of text data to understand and generate human-like language. LLMs can perform various natural language processing (NLP) tasks, including text generation, translation, summarization, and more.
- A key factor in how LLMs work is the way they represent words. Earlier forms of machine learning used a numerical table to represent each word. But this form of representation could not recognize relationships between words such as words with similar meanings. This limitation was overcome by using multi-dimensional vectors, commonly referred to as word embeddings, to represent words so that words with similar contextual meanings or other relationships are close to each other in the vector space.
- Using word embeddings, transformers can pre-process text as numerical representations through the encoder and understand the context of words and phrases with similar meanings as well as other relationships between words such as parts of speech. It is then possible for LLMs to apply this knowledge of the language through the decoder to produce a unique output.

Supply Chain

Definition(s):

1) The term supply chain refers to the linked set of resources and processes between and among multiple levels of an enterprise, each of which is an acquirer that begins with the sourcing of products and services and extends through the product and service life cycle (NIST SP800-161r1-upd1)
2) A linked set of resources and processes between multiple tiers of developers that begins with the sourcing of products and services and extends through the design, development, manufacturing, processing, handling, and delivery of products and services to the acquirer (OMB Circular A-130/CNSSI 4009 dtd 3/2/22).
3) The linked activities associated with providing materiel to end users for consumption. Those activities include supply activities (such as organic and commercial inventory control points (ICPs) and retail supply activities), maintenance activities (such as organic and commercial depot level maintenance facilities and intermediate repair activities), and distribution activities (such as distribution depots and other storage locations, container consolidation points, ports of embarkation and debarkation, and ground, air, and ocean transporters). (DoDI 4140.01)
4) The linked activities associated with providing materiel from a raw material stage to an end user as a finished product in its final disposition. (DAFPD23-1)

Supply Chain Risk

Definition(s):

1) Risks that arise from the loss of confidentiality, integrity, or availability of information or information systems and reflect the potential adverse impacts to organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, other organizations, and the Nation (OMB Circular A-130/CNSSI 4009 dtd 3/2/22).
2) The risk that an adversary may sabotage, maliciously introduce unwanted function, or otherwise subvert the design, integrity, manufacturing, production, distribution, installation, operation, or maintenance of an item of supply or a system so as to surveil, deny, disrupt, or otherwise degrade the function, use, or operation of a system (referenced from The Ike Skelton National Defense Authorization Act for Fiscal Year 2011, Section 806) (CNSSD No. 505/CNSSI 4009 dtd 3/2/22).
3) Supply chain risk is anything that has potential to jeopardize the integrity of products, services, people, and technologies or disrupt the flow of product, materiel, information, and finances across the life cycle of a weapon or support system (DAFI20-101/63-101)

Top 10 Supply Chain Attacks

1. Discord Bot Platform Attack (March 2024) - The Top.gg bot community of Discord, with over 170,000 members, has been impacted by a supply chain attack aimed at infecting developers with malware that steals sensitive information. Over the years, the threat actor has used several tactics, procedures, and techniques, including hijacking GitHub accounts, distributing malicious Python packages, using a fake Python infrastructure, and social engineering. Top.gg was infected by an information-stealing malware after downloading a malicious clone of a tool known as Colorama.

2. Okta Supply Chain Attack (October 2023) - Okta, an authentication and identity management service provider, reported in October 2023 that threat actors could access private consumer data by obtaining credentials to its customer support management system. In recent support cases, the attackers could view files uploaded by specific customers.

3. JetBrains Supply Chain Attack (September/October 2023) - In December, government officials warned that the Solarwinds attackers were exploiting a critical vulnerability in JetBrains TeamCity servers. The critical authentication bypass vulnerability raised attention due to its potential impact and high severity. Unauthenticated intruders with HTTP(S) access can exploit this flaw to gain administrative control of affected servers and execute remote code, presenting a potential vector for supply chain attacks. This attack was carried out by a Russian threat actor named Cozy Bear, who is linked to the Russian Foreign Intelligence Service (SVR RF). In the attack, threat actors gained admin access to the server and employed remote code execution. No user interaction was needed while many large software organizations were using TeamCity servers for their CI/CD, with over 3,000 directly exposed.

Top 10 Supply Chain Attacks

- **MOVEit Supply Chain Attack (June 2023) -** In June, the MOVEit supply chain attack was executed, targeting users of the MOVEit Transfer tool, owned by the US organization Progress Software. MOVEit is designed to transfer sensitive files in a secure manner, and it is popular in the US. The ransomware group Cl0p has been associated with the attack. The attackers used EWIs (Exposed Web Interfaces) to cause significant damage. The web-facing MOVEit app was infected with a web shell called LEMURLOOT, which was then used to steal data from MOVEit transfer databases.
- **3CX Supply Chain Attack (March 2023) -** In March, the 3CX attack targeted macOS and Windows Desktop applications, raising concern about the security and integrity of the software's supply chain. The cyber criminals compromised the application using an infected library file, which subsequently downloaded an encrypted file containing command-and-control information. This enabled the attackers to execute malicious activities within the victim's environment.
- **Microsoft Supply Chain Attack (February 2023) -** In February 2023, a software supply chain attack also affected Microsoft. The attack exploited a vulnerability in the Jfrog Artifactory, a binary repository manager that Microsoft uses to distribute and store its software components. The attackers accessed Jfrog Artifactory and injected malicious code into some of Microsoft's software components, allowing them to access Microsoft's network while stealing source code and other confidential information.

Top 10 Supply Chain Attacks

- **Norton Supply Chain Attack (May 2023) -** Norton's most notable software is its antivirus, which is widely used. They were also attacked in May 2023. The attack used a zero-day vulnerability in MOVEit transfer, an MFT(Managed File Transfer) software that Norton's parent company utilizes to transfer files between consumers and offices. The attackers accessed Norton's network and stole employees' personal information and specific details. The attackers also threatened to release the stolen data if Norton didn't pay a ransom.
- **Airbus Supply Chain Attack (January 2023) -** Airbus was also attacked in January 2023 by a threat actor known as USDoD. The organization confirmed that the attack had been carried out through a compromised employee account at Turkish Airlines, one of Airbus's consumers. The threat actor could access the employee's account and gain access to Airbus's systems. The data breached included personal information associated with over 3000 Airbus vendors, such as Rockwell Collins and the Thales Group. The data dump included names, phone numbers, and email addresses.
- **SolarWinds (Late 2020) -** In late 2020, SolarWinds provided software that contained malware that was intended together with sensitive information wherever it was installed. Customers had complete confidence in the signed software they received, and they believed that it was free of malicious code and viruses as it had not been modified since SolarWinds signed, built, and delivered it to them. However, attackers placed the Sunspot malware into the Orion IT monitoring system and management software utilized by SolarWinds. SolarWinds digitally signed the resultant, which was then utilized to infiltrate over 18,000 private commercial consumers and the government. The malware gathered information from the infected networks and sent data to a remote server. Cozy Bear was again responsible for this attack, which is connected to the Russian Foreign Intelligence Service (SVR).

Top 10 Supply Chain Attacks

- **ShadowHammer/ASUS (2019) -** In 2019, Taiwanese computer manufacturers fell victim to attackers who found critical code signing keys on their web update server. The intruders added malware to legitimate ASUS updates, signed with ASUS's code signing keys, infecting 1 million ASUS computers. The ShadowHammer attacks happened over a period of 6 months. They impacted ASUS notebook customers who enabled the Live Update feature, a utility that automatically searches for and installs new firmware and software updates from ASUS.

Predictive AI Attacks Taxonomy

– Availability Violations (ID: NISTAML.01)
  - Model Poisoning (ID: NISTAML.011)
  - Clean-label Poisoning (ID: NISTAML.012)
  - Data Poisoning (ID: NISTAML.013)
  - Energy-latency (ID: NISTAML.014)

– Integrity Violations (ID: NISTAML.02)
  - Clean-label Poisoning (ID: NISTAML.012)
  - Clean-label Backdoor (ID: NISTAML.021)
  - Evasion (ID: NISTAML.022)
  - Backdoor Poisoning (ID: NISTAML.023)
  - Targeted Poisoning (ID: NISTAML.024)
  - Black-box Evasion (ID: NISTAML.025)
  - Model Poisoning (ID: NISTAML.026)

– Privacy Compromises (ID: NISTAML.03)
  - Model Extraction (ID: NISTAML.031)
  - Reconstruction (ID: NISTAML.032)
  - Membership Inference (ID: NISTAML.033)
  - Property Inference (ID: NISTAML.034)

– Supply Chain Attacks (ID: NISTAML.05)
  - * Model Poisoning (ID: NISTAML.05

Generative AI Attacks Taxonomy

– Availability Violations (ID: NISTAML.01)
  Data Poisoning (ID: NISTAML.013)
  Indirect Prompt Injection (ID: NISTAML.015)
  Prompt Injection (ID: NISTAML.018)
  Integrity Violations (ID: NISTAML.02)
  Data Poisoning (ID: NISTAML.013)
  Indirect Prompt Injection (ID: NISTAML.015)
  Prompt Injection (ID: NISTAML.018)
  Backdoor Poisoning (ID: NISTAML.023)
  Targeted Poisoning (ID: NISTAML.024)
  Misaligned Outputs (ID: NISTAML.02
– Misuse Violations (ID: NISTAML.04)
  * Prompt Injection (ID: NISTAML.018)
– Supply Chain Attacks (ID: NISTAML.05)
  * Model Poisoning (ID: NISTAML.05
– Privacy Compromises (ID: NISTAML.03)
  Indirect Prompt Injection (ID: NISTAML.015)
  Prompt Injection (ID: NISTAML.018)
  Backdoor Poisoning (ID: NISTAML.023)
  Membership Inference (ID: NISTAML.033)
  Prompt Extraction (ID: NISTAML.035)
  Leaking information from user interactions (ID: NISTAML.036)
  Training Data Attacks (ID: NISTAML.037)
  Data Extraction (ID: NISTAML.038)
  Compromising connected resources (ID: NISTAML.03

Attack Classification

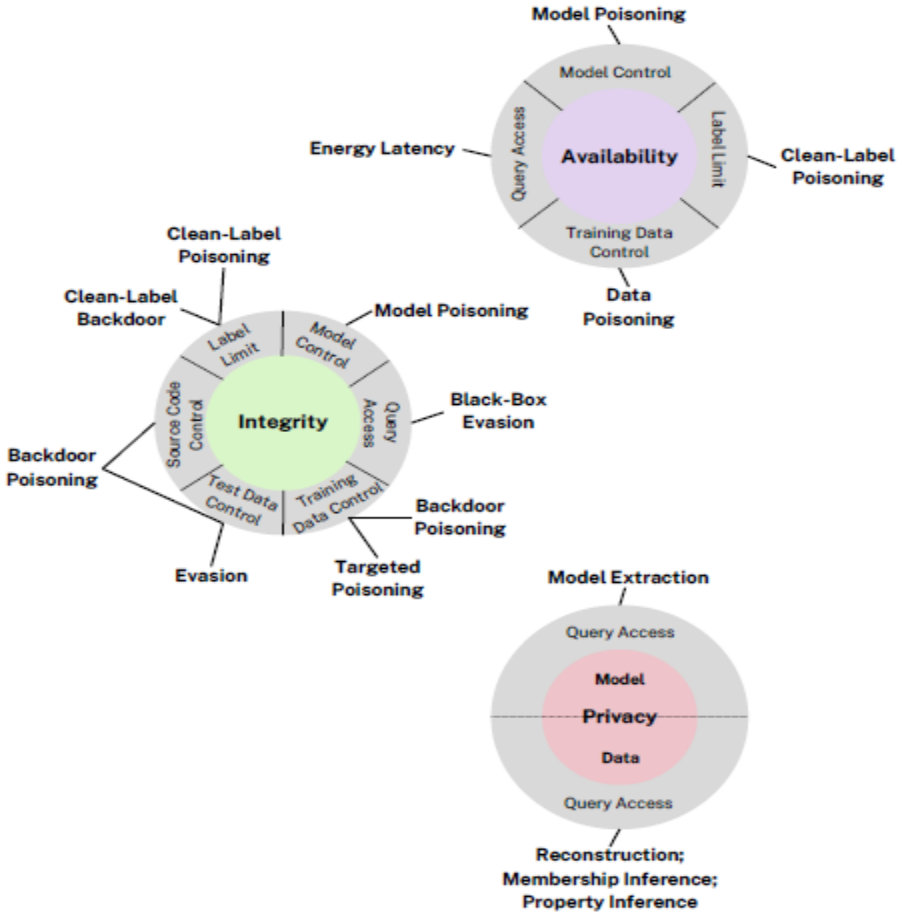


Figure 1. Taxonomy of attacks on PredAI systems

The attacker's objectives are shown as disjointed circles with the attacker's goal at the center of each circle: **availability** breakdown, **integrity** violation, and **privacy** compromise. The capabilities that an adversary must leverage to achieve their objectives are shown in the outer layer of the objective circles. Attack classes are shown as callouts connected to the capabilities required to mount each attack. Multiple attack classes that require the same capabilities to reach the same objective are shown in a single callout.

These attacks are classified according to the following dimensions: 1) learning method and stage of the learning process when the attack is mounted, 2) attacker goals and objectives, 3) attacker capabilities, and 4) attacker knowledge of the learning process. Several adversarial attack classification frameworks have been introduced in prior works, and the goal here is to create a standard terminology for adversarial attacks on ML that unifies existing work.

Cybersecurity LLM Attacks

The first poisoning attacks were discovered in cybersecurity for worm signature generation (2006) and spam email classification (2008). Since then, poisoning attacks have been shown for malware classification, malicious PDF detection, and Android malicious app classification. Evasion attacks against similar data modalities have been proposed as well: malware classification, PDF malware classification, Android malicious app detection, and network intrusion detection. Poisoning unsupervised learning models has been shown for clustering used in malware classification and network traffic anomaly detection.

Anomaly detection based on data-centric approaches allows for automated feature learning through ML algorithms. However, the application of ML to such problems comes with specific challenges related to the need for very low false negative and low false positive rates (e.g., the ability to catch zero-day attacks). This challenge is compounded by the fact that trying to accommodate all these together makes ML models susceptible to adversarial attacks.

- **Tabular data:** There have been numerous attacks against ML models working on tabular data, such as poisoning availability attacks against healthcare and business applications, privacy attacks against healthcare data, and evasion attacks against financial applications.

Recently, the use of ML models trained on multimodal data has gained traction, particularly the combination of image and text data modalities. Several papers have shown that multimodal models may provide some resilience against attacks, but other papers show that multimodal models themselves could be vulnerable to attacks mounted on all modalities at the same time.

GenAI Attacks

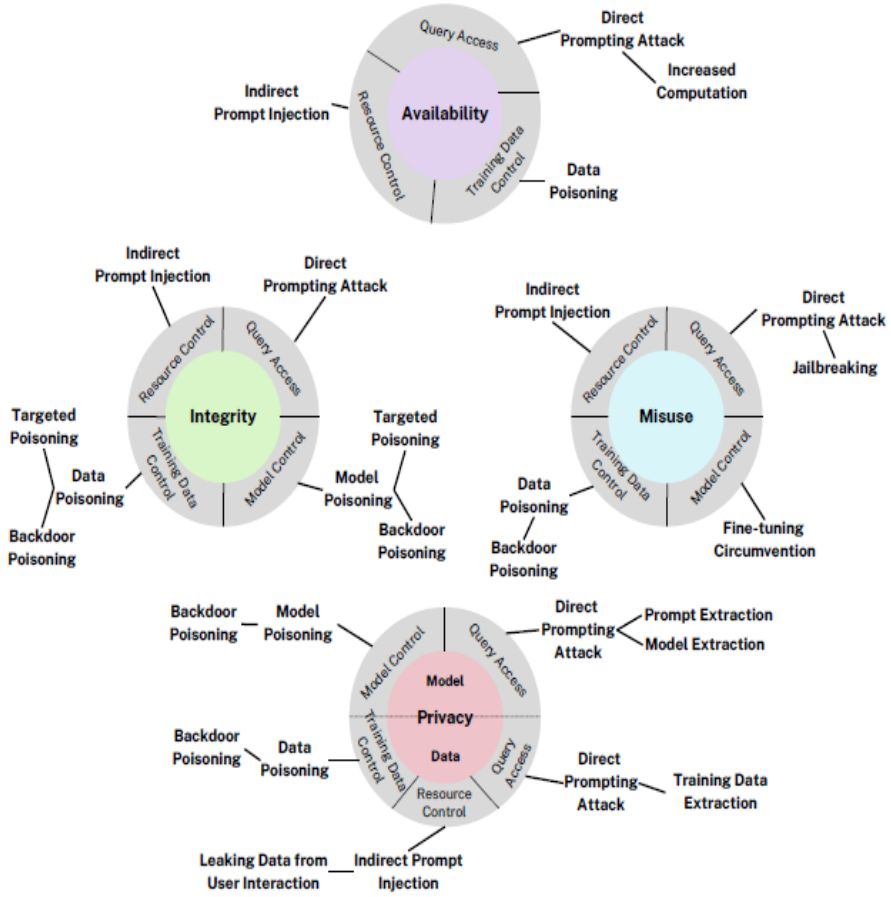


Figure 2. Taxonomy of attacks on GenAI systems

3.1.1. GenAI Stages of Learning

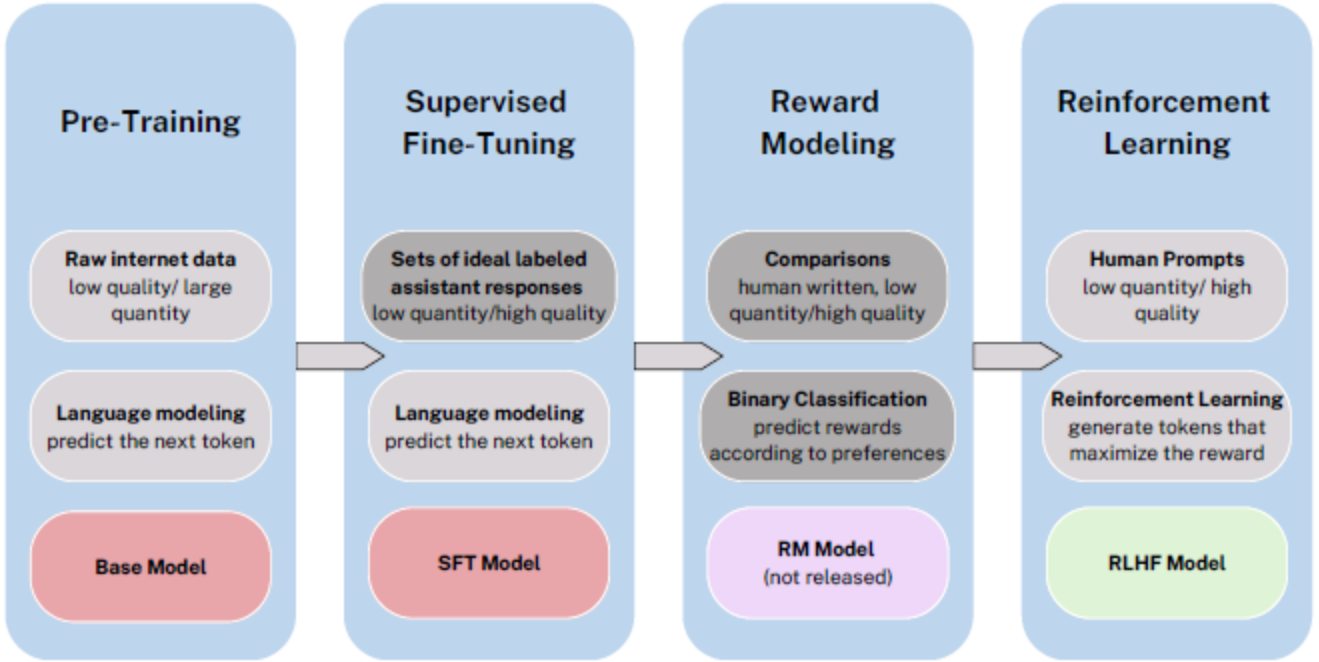


Figure 3. Example LLM Training Pipeline used for InstructGPT [281]

Supply Chain Attacks and Mitigations

Since AI is software, it inherits many of the vulnerabilities of the traditional software supply chain, such as reliance on third-party dependencies. AI development also introduces new types of dependencies, including data collection and scoring, the integration or adaptation of third-party-developed AI models, and the integration of third-party-developed plugins into AI systems.

Mitigating the security challenges in AI supply chain management is complex and requires a multifaceted approach that combines existing practices for software supply chain risk management with the management of AI-specific supply chain risks, such as using provenance information for the additional artifacts involved.

Studies of real-world security vulnerabilities against ML suggest that security is best addressed comprehensively and by considering the full attack surface, including data and model supply chains, software, and network and storage systems.

While all these supply chain risks are critical in the broader context of securing AI systems, there are certain types of attacks that rely on exploiting the specific statistical and data-based properties of ML systems, thus falling within the domain of AML.

Data Poisoning Attack

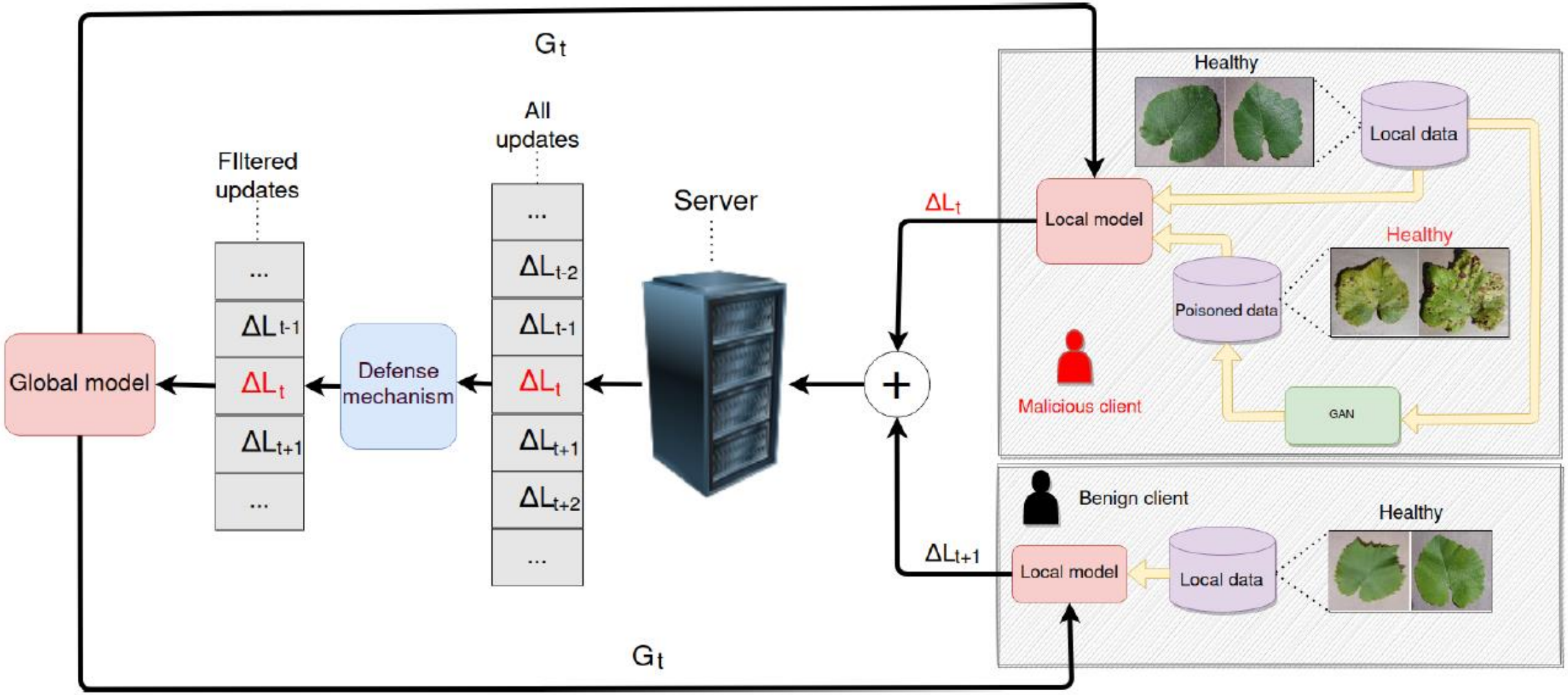
The performance of GenAI text-to-image and text-to-text models has been found to scale with dataset size (among other properties like model size and data quality); for example, Hoffmann et al. suggest compute-optimally training a 520 billion parameter model may require 11 trillion tokens of training data. Thus, it has become common for GenAI foundation model developers to scrape data from a wide range of sources. In turn, the scale of this data and the diversity of its sources provides a large potential attack surface into which attackers may seek to insert adversarial constructed data points. For example, dataset publishers may provide a list of URLs to constitute a training dataset, and attackers may be able to purchase some of the domains that serve those URLs and replace the site content with their own malicious content.

Beyond the vast quantities of pre-training data, data poisoning attacks may also affect other stages of the LLM training pipeline, including instruction tuning and reinforcement learning from human feedback, which may intentionally source data from many human participants.

As with PredAI models, data poisoning attacks could lead to attackers controlling model behavior through the insertion of a backdoor such as a word or phrase that, when submitted to a model, acts as a universal JAILBREAK. Attackers could also use data poisoning attacks to modify model behavior on user queries, such as causing the model to incorrectly summarize or otherwise produce degenerate outputs in response to queries that contain a particular trigger word or phrase. These attacks may be practical—requiring a relatively small portion of the total dataset—and may lead to a range of bad outcomes, such as code suggestion models which intentionally suggest insecure code.

Example of Data Poisoning Attack

Psychogyios, K.; Velivassaki, T.-H.; Bourou, S.; Voulkidis, A.; Skias, D.; Zahariadis, T. GAN-Driven Data Poisoning Attacks and Their Mitigation in Federated Learning Systems. Electronics **2023**, 12, 1805. https://doi.org/10.3390/electronics12081805



Around the Room