

A blue-tinted photograph of two men in an office setting. An older man with a grey beard and glasses is on the left, looking towards a laptop. A younger man with a beard and glasses is on the right, looking at the laptop. The background shows office shelves with various papers and sticky notes. The word 'HIGHPOINT' is in the top left corner.

HIGHPOINT

2025 Secure AI Whitepaper

info@highpoint.com | HighPoint.com

Table of Contents

03 Executive Summary

04 AI FAQs

07 AI Security Risks

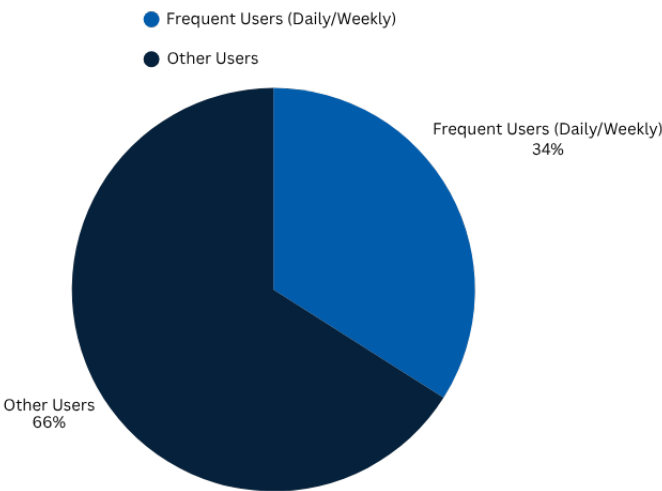
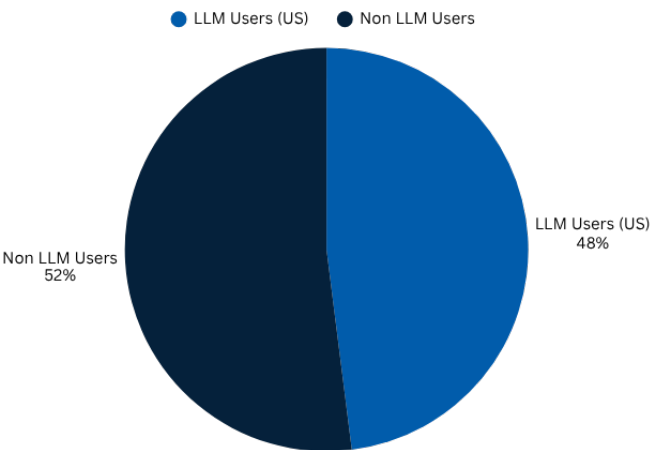
09 Best Practices for the secure
usage of AI

Executive Summary

Have you used AI to get a quick answer on statistics for use in a work report or to compare stats of your favorite sports players? If so, you are certainly not alone. Elon University reports that as of early 2025, **around half (52%) of U.S. adults** use AI Large Language Models (LLMs) like ChatGPT, Gemini, Claude, and Copilot. Of those LLM users, about 34% use them at least once a day or several times a week for personal, informal learning, and work activities. As of early to mid-2025, approximately 78% of companies worldwide have adopted AI technologies in at least one business function. This is a significant jump from 55% in 2023.

With the widespread use of AI, as with the adoption of any new technology, caution must be taken to ensure informed and secure usage. While the security principles around AI are not new, it is of vital importance to ensure that best practices are followed when using AI. This is especially true when introducing sensitive data into LMMs for both public and private (self-developed) models.

HighPoint recommends solutions to ensure the secure usage of AI, for use with both publicly-available and private, self-developed AI capabilities. We are happy to discuss AI security in your environment and consult with you to ensure that your usage of AI is secure, keeping sensitive data private and ensuring that your AI solutions are not used for nefarious purposes.



AI FAQ's

What exactly is AI?

Artificial Intelligence (AI) is a broad field of computer science that focuses on creating machines or systems that can simulate or mimic human intelligence. The goal of AI is to enable computers to perform tasks that typically require human cognitive abilities.

Key aspects of what constitutes AI include:

- **Learning:** AI systems can learn from data, identifying patterns and relationships without being explicitly programmed for every scenario. This is often achieved through Machine Learning (ML), a subset of AI where algorithms are trained on vast datasets.
- **Reasoning:** AI can process information and apply logical rules or inferred patterns to make decisions or draw conclusions.
- **Problem-solving:** AI systems can analyze problems and devise strategies to achieve desired outcomes.
- **Perception:** This involves the ability to interpret sensory data, such as images (Computer Vision) or sounds (Speech Recognition), to understand the environment.
- **Language Understanding and Generation (Natural Language Processing – NLP):** AI can comprehend, interpret, and generate human language, enabling communication with humans in natural ways (e.g., chatbots, language translation).
- **Decision-making:** Based on learned patterns and reasoning, AI can make choices or recommendations.
- **Adaptability:** Many AI systems can improve their performance over time as they are exposed to more data and feedback.

Important Distinctions:

- **Narrow AI (or Weak AI):** This is the vast majority of AI in use today. It's designed and trained for specific tasks (e.g., facial recognition, playing chess, recommending products, self-driving cars). It can perform these tasks very well, often surpassing human ability, but it lacks broader human-like intelligence.
- **Artificial General Intelligence (AGI or Strong AI):** This is a theoretical concept referring to AI that possesses human-level cognitive abilities across a wide range of tasks, capable of learning, reasoning, and adapting to any intellectual task a human can. AGI does not currently exist.
- **Artificial Superintelligence (ASI):** Another theoretical concept, where AI surpasses human intelligence in virtually every field, including creativity, general knowledge, and problem-solving. This is even further beyond AGI.

In essence, AI aims to create “intelligent” machines that can automate complex processes, assist humans in various tasks, and even discover new insights from vast amounts of data.

I've heard the term "LLM" thrown around quite a bit, but what is it?

An **LLM** stands for **Large Language Model**.

At its core, an LLM is a type of artificial intelligence program designed to understand and generate human-like text. It's called "large" because these models are trained on absolutely massive datasets of text and code – billions, even trillions, of words from books, articles, websites, conversations, and more. This vast amount of training data allows them to learn complex patterns, grammar, semantics, and even some aspects of world knowledge.

How they work (simplified):

Imagine feeding billions of sentences into a system. The system tries to guess the next word in each sentence. When it's wrong, it adjusts its internal parameters (billions of them) to try and get closer to the correct answer next time. Over countless iterations and with immense computational power, these adjustments allow the model to learn the intricate statistical relationships between words and concepts.

Common Applications:

- **Content Generation:** Writing articles, stories, marketing copy, social media posts.
- **Chatbots and Virtual Assistants:** Powering conversational AI for customer service, information retrieval, and interactive experiences.
- **Summarization:** Condensing long texts into shorter, digestible summaries.
- **Translation:** Translating text between different languages.
- **Code Generation:** Writing code snippets, debugging, or explaining code.
- **Question Answering:** Providing direct answers to user queries.
- **Brainstorming and Ideation:** Helping users explore different ideas and perspectives.

Popular examples of LLMs include models like Google's Gemini, OpenAI's GPT series, Anthropic's Claude, and Meta's Llama.

What is an AI Agent?

An AI agent is a software program or system that can perceive its environment, process that information, make decisions, and then take actions autonomously or semi-autonomously to achieve specific goals.

Think of it as a more advanced form of AI than a simple chatbot or a model that just generates text. While a Large Language Model (LLM) might be the "brain" of an AI agent, the agent itself has the capacity to:

- **Perceive:** Gather information from its environment. This can be anything from text input, sensor data (for a robot), web pages, or even data from other software systems.
- **Reason/Plan:** Process the gathered information, understand the given goal, and develop a strategic plan or sequence of steps to achieve that goal. This often involves using an LLM for understanding and generating ideas, and then breaking down complex tasks into smaller, manageable subtasks.

- **Act:** Take actions in its environment. This could be generating a response, sending an email, interacting with a website, controlling a robot, or updating a database. AI agents often use “tools” (like web browsers, APIs, or specific software functions) to interact with the real world or other digital systems.
- **Learn/Adapt:** Improve its performance over time by learning from its experiences, feedback, and new information. It can remember past interactions and adjust its future decisions.
- **Autonomy:** This is a key differentiator. While AI assistants might respond to direct user prompts, AI agents can initiate actions, make decisions independently, and adapt to changing conditions to proactively pursue their objectives with minimal human intervention.

Key Differences from other AI concepts:

- **AI Assistants (like Siri or Alexa):** While helpful, assistants are generally more reactive. They respond to direct commands and provide information or complete simple tasks, with the user typically making the final decisions. AI agents have a higher degree of autonomy and can proactively work towards a goal.
- **Bots:** Bots are typically the least autonomous. They follow pre-defined rules and are designed for very specific, often repetitive, tasks or basic conversations. They have limited or no learning capabilities.

Examples of AI Agents:

- **Customer Service Agents**
- **Project Management Agents**
- **Financial Trading Agents**
- **Self-Driving Cars**
- **Software Development Agents**

In essence, AI agents are moving towards a paradigm where AI systems are not just tools that react to specific inputs, but proactive, goal-oriented entities that can independently navigate complex situations and achieve objectives.

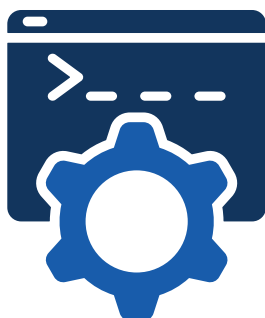


AI Security Risks

As of mid-2025, the security of Artificial Intelligence (AI) systems has become a paramount concern for organizations worldwide. The ranking of these risks is determined by a combination of their prevalence, potential for direct impact, and the unique challenges they present to both new and existing security paradigms. Including the critical risk of data leakage, here are the top five risks in AI security, ranked in order of importance.

01. Prompt Injection & Malicious Input

Why it's #1: This remains the most immediate and widespread threat, particularly with the proliferation of Large Language Models (LLMs). It is a direct attack on the primary interface of generative AI. By crafting malicious inputs (prompts), an attacker can bypass the model's safety filters and original instructions. This can be used to generate harmful content, execute unauthorized functions, or, most critically, serve as a vector for the number two risk: data leakage. Its top ranking is due to the low barrier to entry for attackers and the direct, immediate control it can yield over an AI agent's behavior.



02. Sensitive Information Disclosure (Data Leakage)

Why it's #2: Data leakage is a severe and high-impact risk where an AI model inadvertently reveals confidential information it was exposed to during training or has access to in its operational context. This can include trade secrets, intellectual property, personal identifiable information (PII), or system credentials. As confirmed by its high placement in frameworks like the OWASP Top 10 for LLMs, this risk is not just theoretical; it's a fundamental vulnerability. An attacker can use carefully designed prompts—often through prompt injection—to coax the model into revealing this sensitive data, leading to massive privacy breaches, regulatory fines, and loss of competitive advantage.

03. Data Poisoning

Why it's #3: Data poisoning is an insidious attack that corrupts the integrity of an AI model at its core. By injecting manipulated or malicious data into the training set, attackers can create hidden backdoors, introduce biases, or degrade the model's performance in subtle but critical ways. This threat is ranked highly because it is difficult to detect, and its effects can be long-lasting and catastrophic. A compromised model may make consistently flawed decisions, from misdiagnosing medical conditions to approving fraudulent financial transactions. The remedy often requires a complete and expensive retraining of the model from a clean dataset.

04. AI Supply Chain Vulnerabilities

Why it's #4: Modern AI systems are rarely built in isolation. They are assembled from a complex supply chain that includes third-party datasets, pre-trained models from public repositories (like Hugging Face), and various software libraries. A vulnerability in any link of this chain can compromise every system that uses it. An attacker could upload a poisoned model to a public hub, which is then unknowingly integrated by thousands of developers into their applications. This makes AI supply chain attacks a potent vector for widespread disruption and places a significant burden on organizations to vet every component of their AI systems.

5. AI-Powered Social Engineering & Disinformation

Why it's #5: This risk involves the use of AI as a tool to dramatically scale and enhance the effectiveness of attacks on humans. Generative AI can create hyper-realistic deepfake videos and audio, craft highly personalized and convincing phishing emails, and automate the spread of disinformation on a massive scale. This erodes trust in digital communication and can be used to manipulate public opinion, defraud individuals, or impersonate executives to authorize fraudulent transactions. This threat is critically important because it targets the weakest link in many security systems—the human element—with unprecedented sophistication and scale.



Best Practices for the Secure Usage of AI

Grounded in basic security principles, here are the top 5 security best practices when using AI:

1. Implement Robust Data Governance and Privacy Controls:

- **Practice:** This goes beyond mere compliance and involves a proactive approach to managing the data that fuels your AI.
- **How:**
 - **Strict Access Controls:** Apply the principle of least privilege, ensuring that only authorized personnel and systems have access to AI training data, models, and outputs. This includes granular role-based access control (RBAC).
 - **Data Minimization:** Only collect, store, and use the data that is absolutely necessary for the AI's purpose.
 - **Anonymization/Pseudonymization:** Whenever possible, use techniques like data masking, tokenization, or differential privacy to de-identify sensitive information, especially in training and testing datasets.
 - **Encryption:** Encrypt data at rest (storage), in transit (network communication), and even during processing where feasible (e.g., homomorphic encryption for highly sensitive computations, though this is still an emerging field).
 - **Data Provenance and Integrity:** Track the origin and modifications

of all data used by AI systems. Implement checksums and cryptographic hashes to verify data integrity and prevent tampering or poisoning of training data.

2. Secure the Entire AI/ML Development Lifecycle (MLSecOps):

- **Practice:** Integrate security into every phase of AI development, from data collection and model training to deployment and monitoring.
- **How:**
 - **Secure Development Practices:** Apply secure coding principles to AI model development, including vulnerability scanning of libraries and frameworks.
 - **Version Control and Auditing:** Maintain strict version control for models, code, and data. Ensure all changes are logged and auditable.
 - **Supply Chain Security:** Vet third-party AI models, libraries, and components for vulnerabilities. Be aware of the risks associated with open-source AI.
 - **Secure Deployment:** Deploy AI models in secure, isolated environments with robust network segmentation, firewalls, and secure configurations.
 - **Automated Security Testing:** Incorporate automated security testing throughout the CI/CD pipeline for AI systems.

3. Perform Adversarial Testing and Model Hardening:

- Practice: Actively seek out and mitigate vulnerabilities specific to AI models, such as adversarial attacks and model manipulation.
- How:
 - **Adversarial Training:** Train models with adversarial examples to make them more robust against subtle, malicious inputs designed to trick the AI.
 - **Input Validation and Sanitization:** Implement strict validation and sanitization of all inputs to AI models to prevent prompt injection, data manipulation, and other forms of adversarial attacks.
 - **Regular Red Teaming:** Conduct dedicated "red team" exercises where security experts simulate attacks against your AI systems to uncover vulnerabilities before malicious actors do. This includes evasion, poisoning, and model extraction attacks.
 - **Anomaly Detection:** Implement systems to detect unusual model behavior or output that could indicate an adversarial attack or model compromise.

4. Establish Comprehensive Monitoring, Logging, and Incident Response:

- Practice: Continuously monitor AI systems for anomalous behavior, security incidents, and performance degradation, and have a clear plan for responding to breaches.
- How:
 - **Continuous Monitoring:** Monitor model performance, data inputs, outputs, and resource usage for any deviations from baseline behavior. This helps detect data drift, model drift, and potential attacks.
 - **Detailed Logging:** Log all significant AI system activities, including data access, model inferences, configuration changes, and errors. These logs are crucial for forensic analysis during incidents.
 - **AI-Specific Incident Response**
 - Plan:** Develop an incident response plan tailored to AI systems, outlining procedures for detecting, containing, eradicating, and recovering from AI-related security incidents.
 - **Alerting:** Set up real-time alerts for suspicious activities, performance anomalies, or detected adversarial inputs.

5. Develop Strong AI Governance and Ethical Guidelines:

- **Practice:** Implement a formal framework for responsible AI use, encompassing ethical considerations, accountability, and regulatory compliance.
- **How:**
 - **Clear Policies and Guidelines:** Establish clear organizational policies for AI development, deployment, and use, addressing security, privacy, fairness, and transparency.
 - **Cross-Functional Teams:** Involve legal, ethical, and security experts alongside AI developers to ensure a holistic approach to AI governance.
 - **Regular Audits and Compliance Checks:** Conduct regular audits to ensure adherence to internal policies and external regulations (e.g., industry-specific AI guidelines, data privacy laws).
 - **Transparency and Explainability (XAI):** Strive for explainable AI where possible, enabling human understanding of AI decision-making. This aids in identifying biases, debugging issues, and proving accountability.
 - **Employee Training and Awareness:** Educate all employees involved in AI about security best practices, potential AI-specific risks, and their role in maintaining a secure AI environment.

