AI Red Teaming

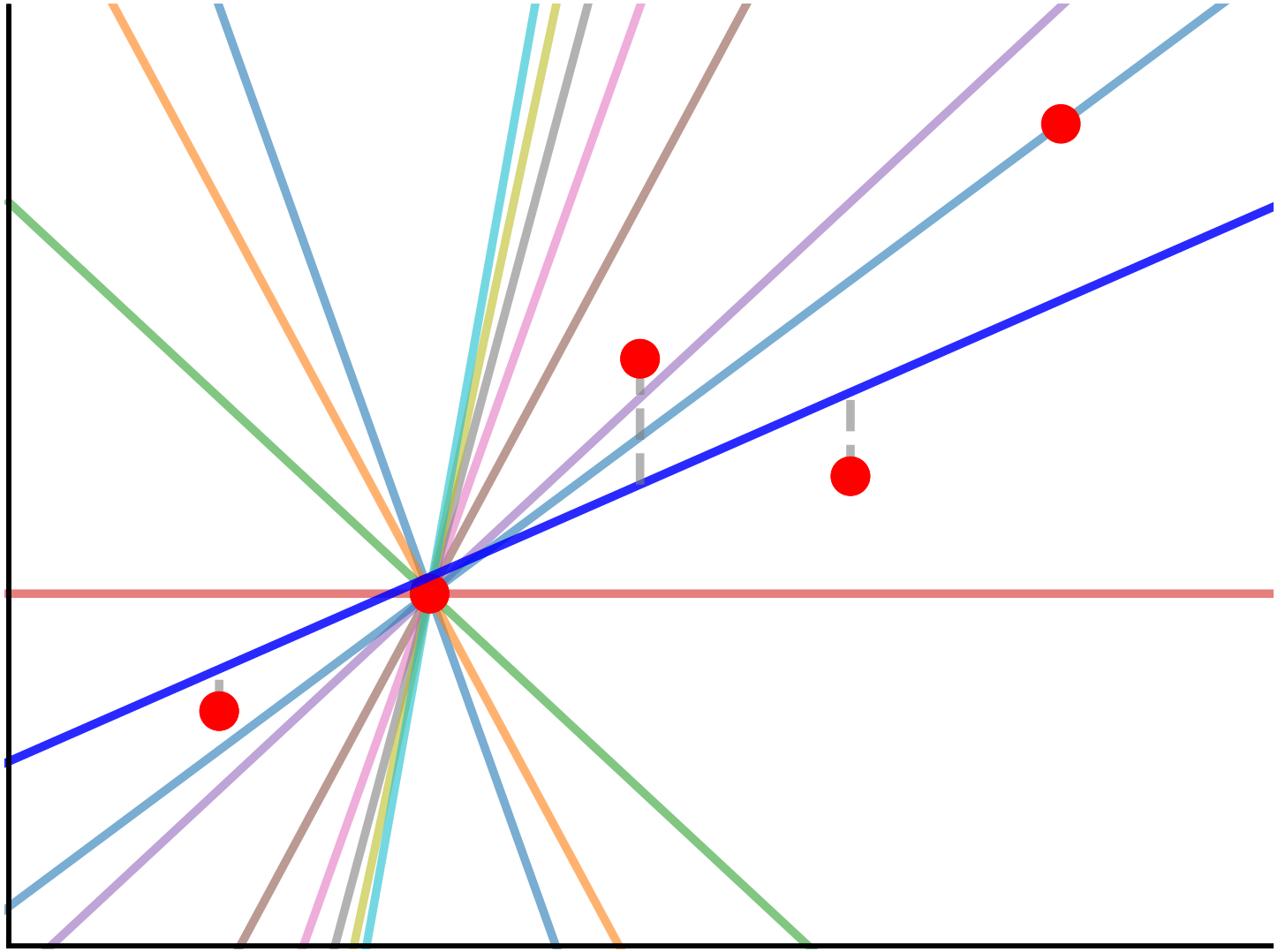
Federico Cerutti

federico.cerutti@unibs.it

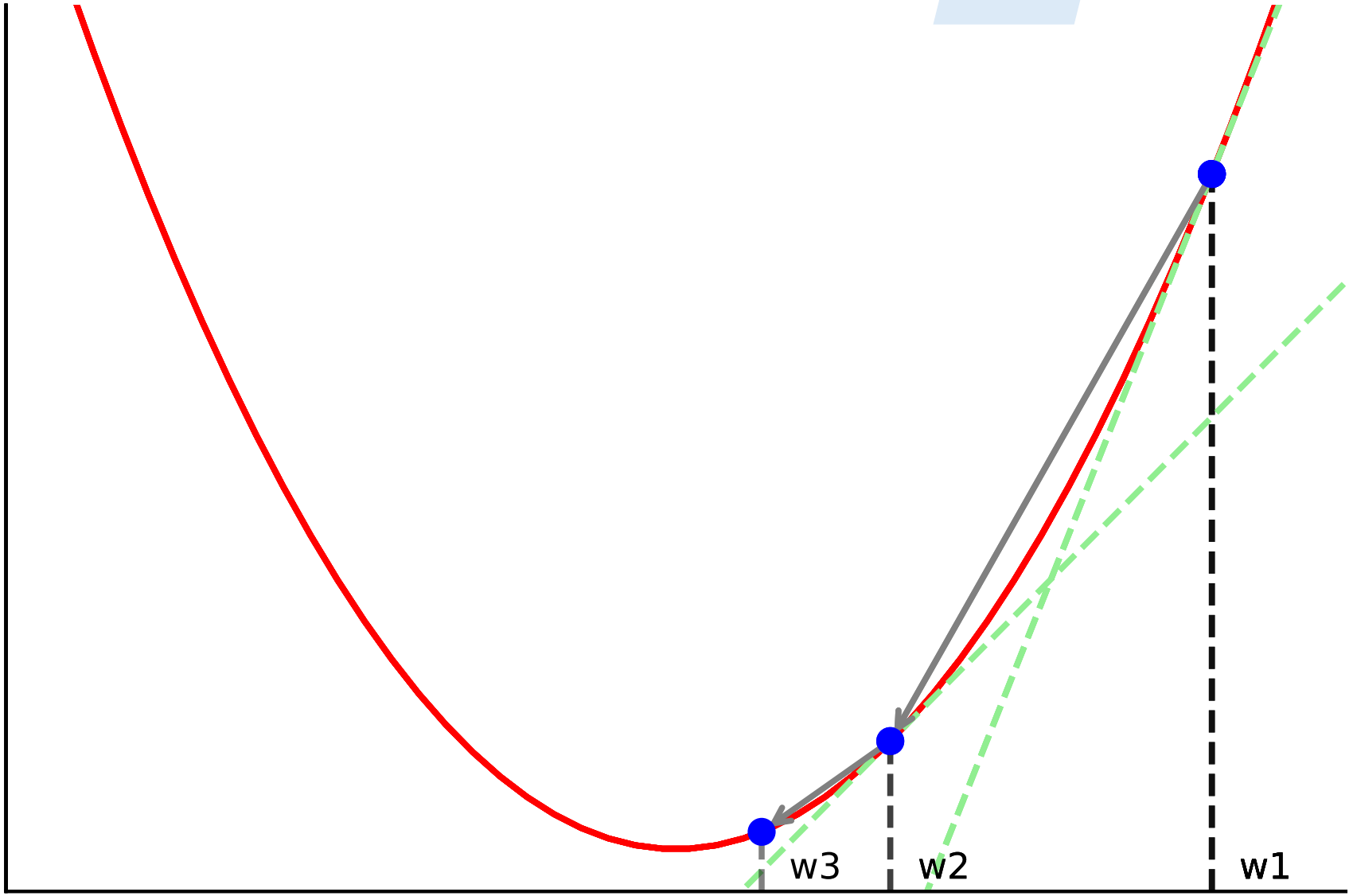
$$\max_{x \in K} \left| f(x) - \sum_{k=1}^{N} a_k h(w_k x + b_k) \right| \leq \varepsilon$$

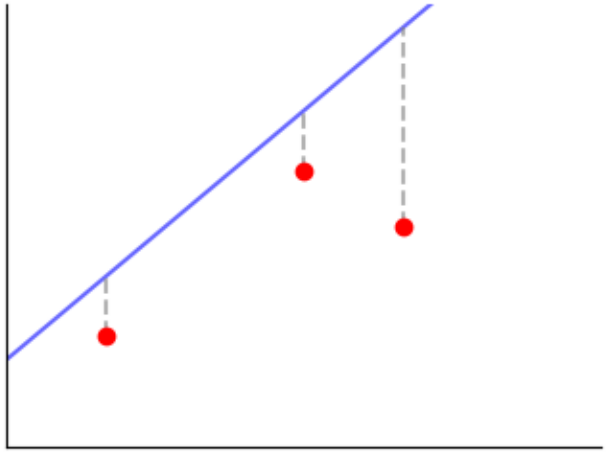
$$y = wx + b$$

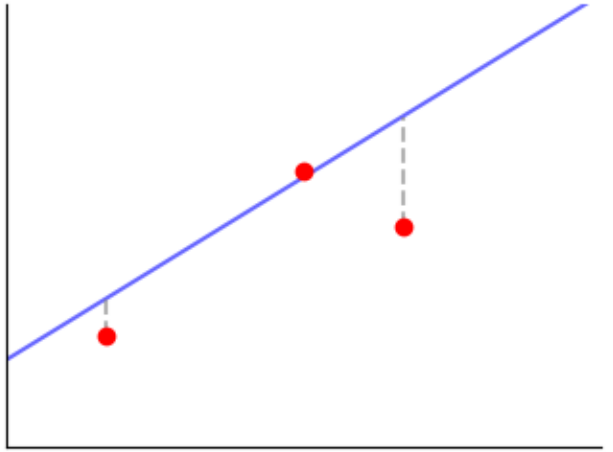
$$y = wx + b$$

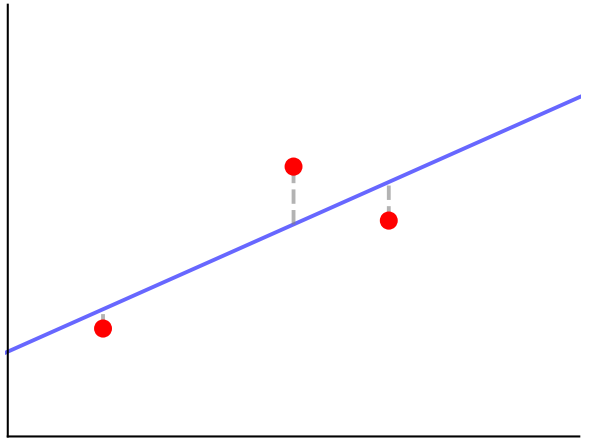


$$y = wx + b$$



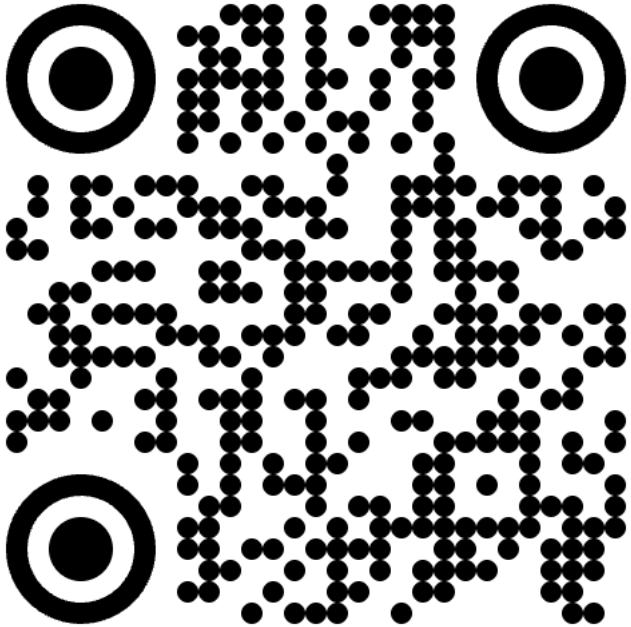






NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)



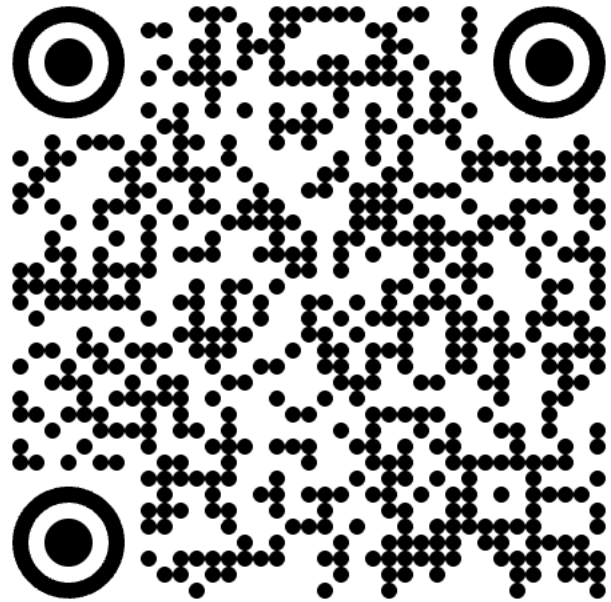
https://www.nist.gov/itl/ai-risk-management-framework


Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

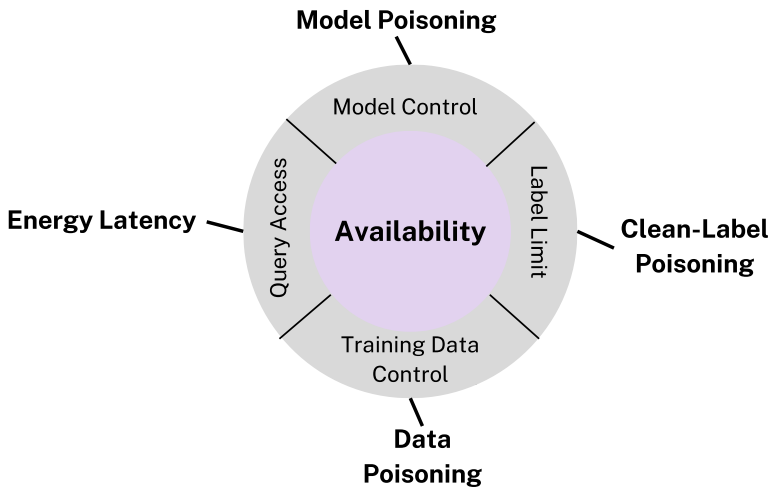
NIST Trustworthy and Responsible AI
NIST AI 100-2e2025

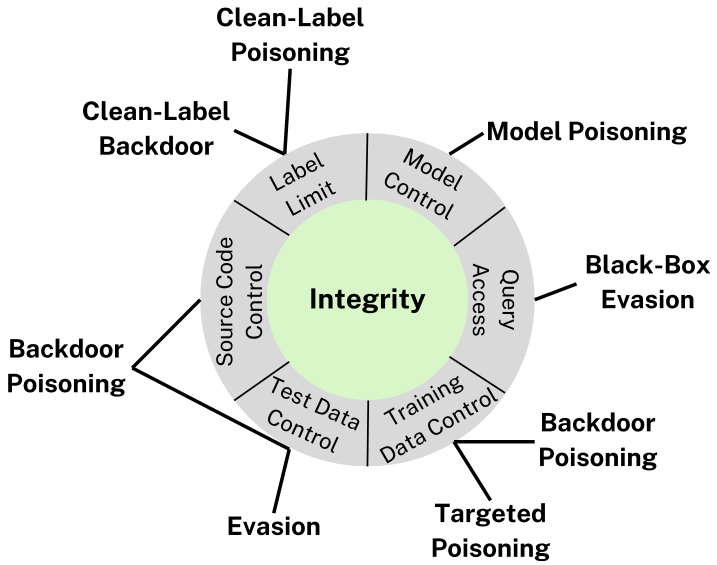
Adversarial Machine Learning

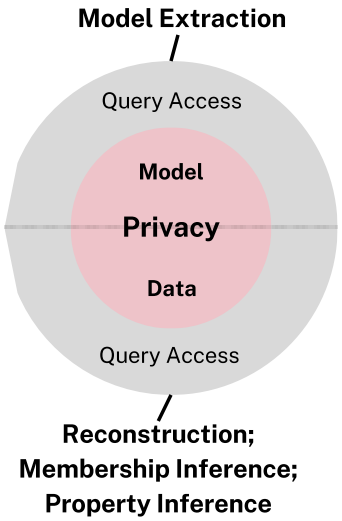
A Taxonomy and Terminology of Attacks and Mitigations

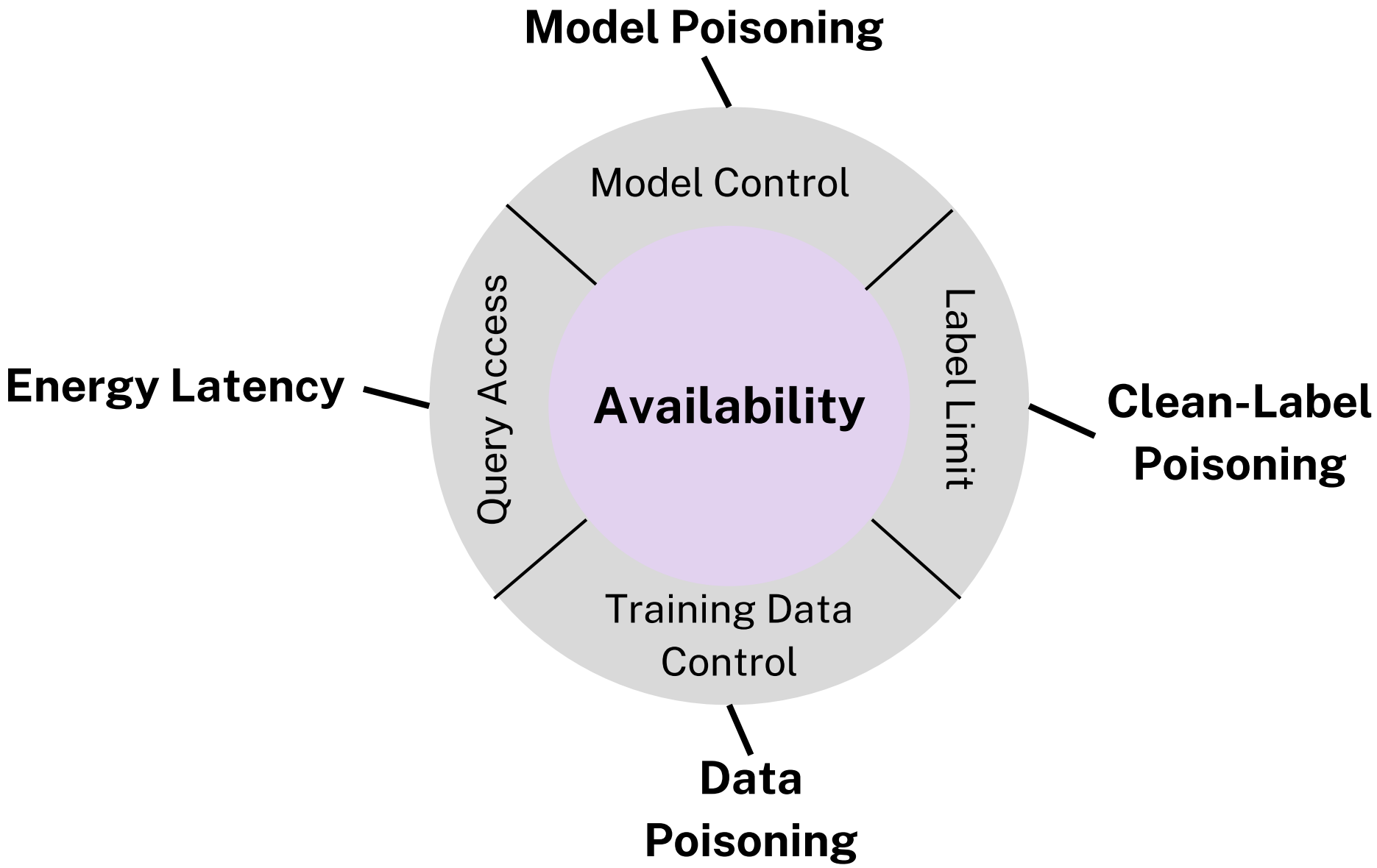


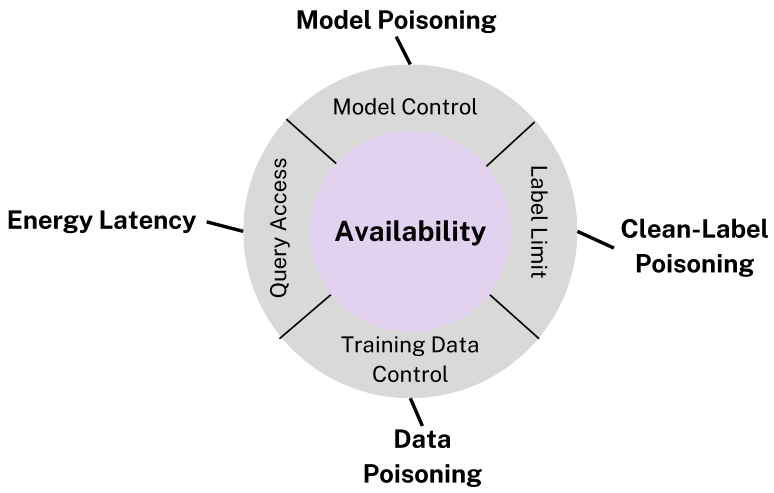
https://csrc.nist.gov/pubs/ai/100/2/e2025/final

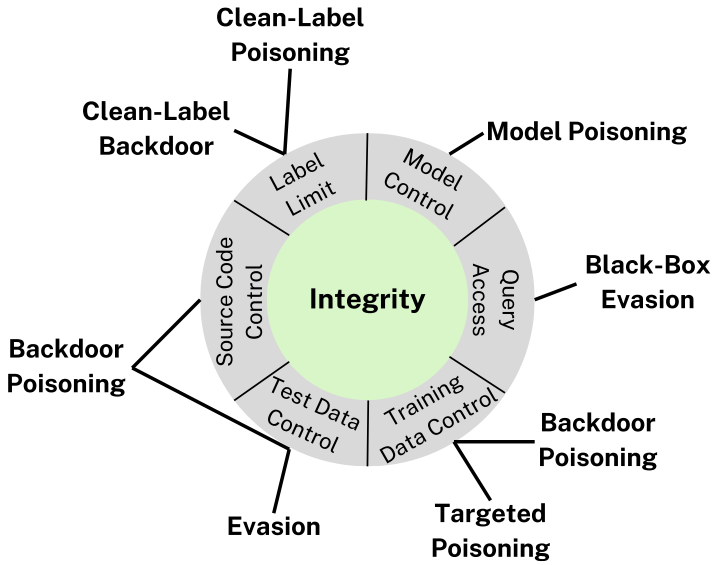

Model Poisoning
Model Control
Energy Latency
Query Access
Availability
Label Limit
Clean-Label Poisoning
Training Data Control
Data Poisoning

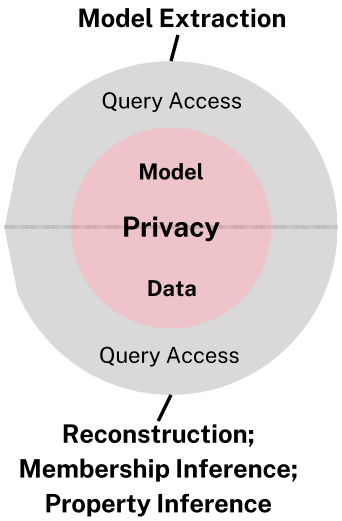

Clean-Label Poisoning
Clean-Label Backdoor
Model Poisoning
Label Limit
Model Control
Source Code Control
Integrity
Query Access
Black-Box Evasion
Backdoor Poisoning
Test Data Control
Training Data Control
Backdoor Poisoning
Evasion
Targeted Poisoning

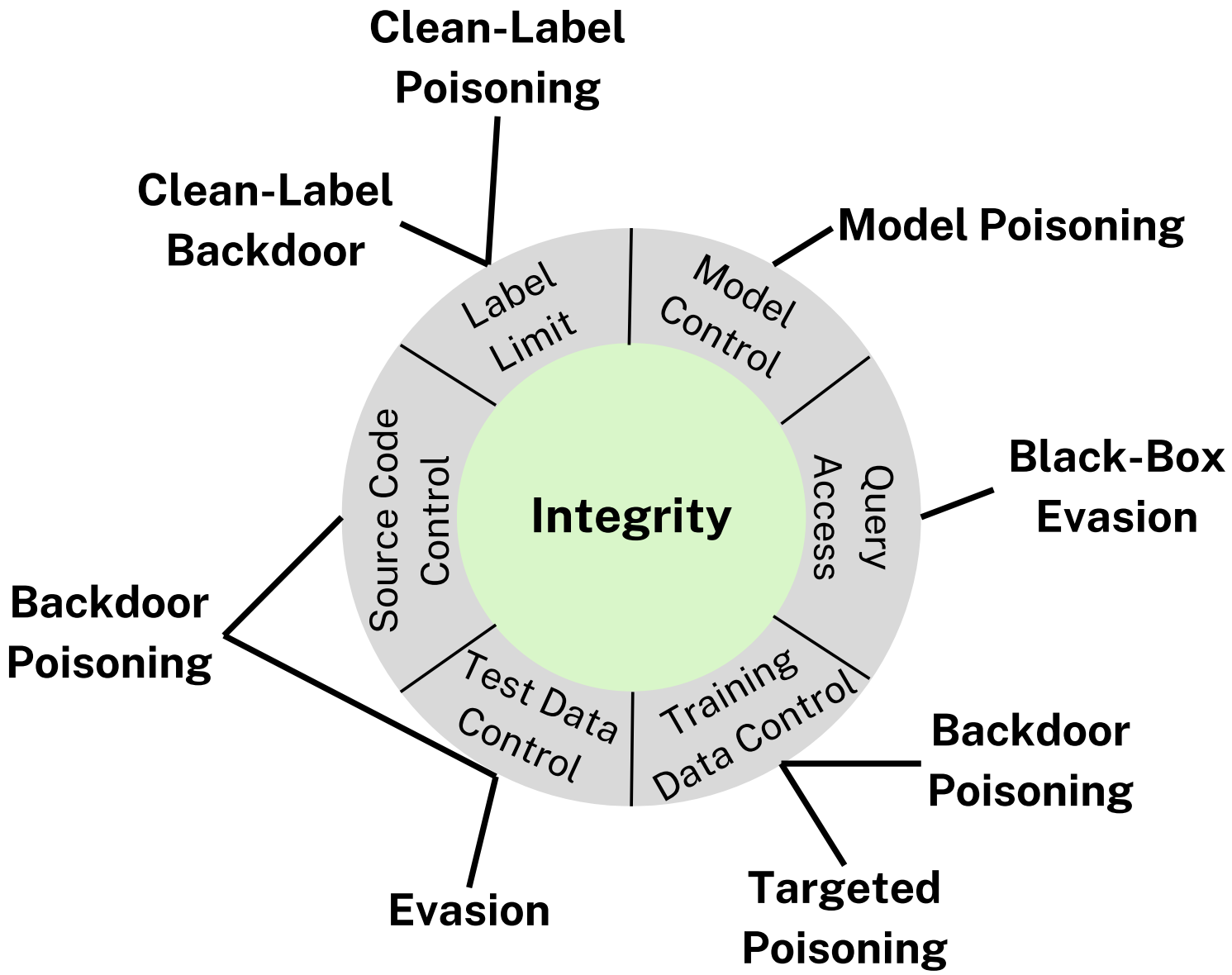

Model Extraction
Query Access
Model
Privacy
Data
Query Access
Reconstruction; Membership Inference; Property Inference

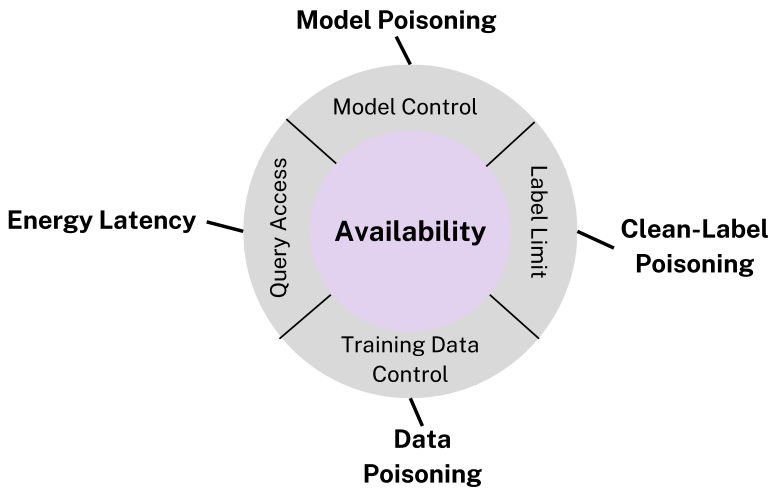

Model Poisoning
Model Control
Query Access
Label Limit
Energy Latency
Availability
Clean-Label Poisoning
Training Data Control
Data Poisoning

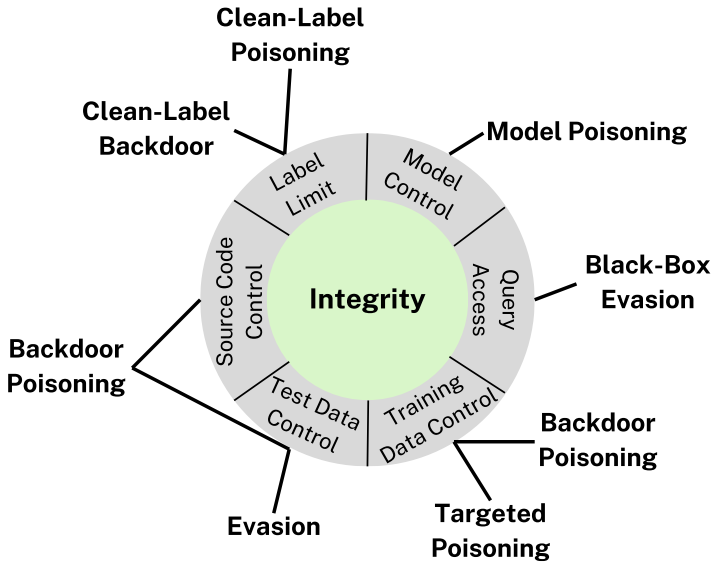

Model Poisoning
Model Control
Energy Latency
Query Access
Availability
Label Limit
Clean-Label Poisoning
Training Data Control
Data Poisoning

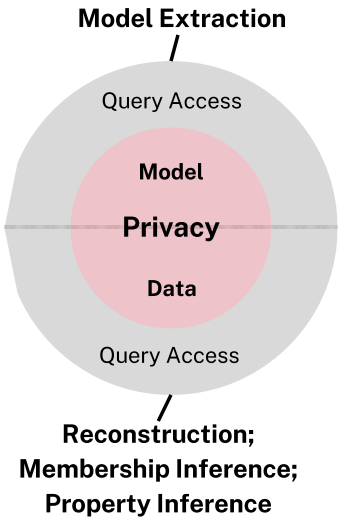

Clean-Label Poisoning
Clean-Label Backdoor
Model Poisoning
Label Limit
Model Control
Source Code Control
Integrity
Query Access
Black-Box Evasion
Backdoor Poisoning
Test Data Control
Training Data Control
Backdoor Poisoning
Evasion
Targeted Poisoning

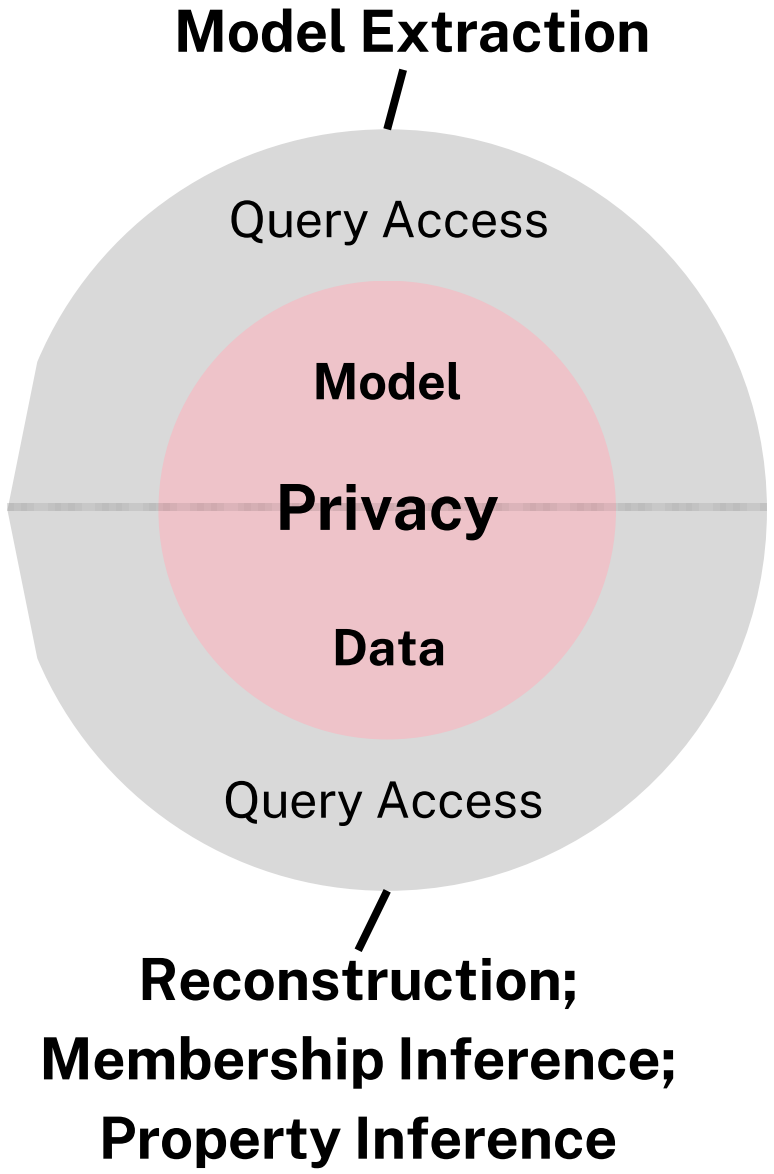

Model Extraction
Query Access
Model
Privacy
Data
Query Access
Reconstruction; Membership Inference; Property Inference


Clean-Label Poisoning
Clean-Label Backdoor
Model Poisoning
Black-Box Evasion
Backdoor Poisoning
Evasion
Backdoor Poisoning
Targeted Poisoning
Label Limit
Model Control
Query Access
Training Data Control
Test Data Control
Source Code Control
Integrity

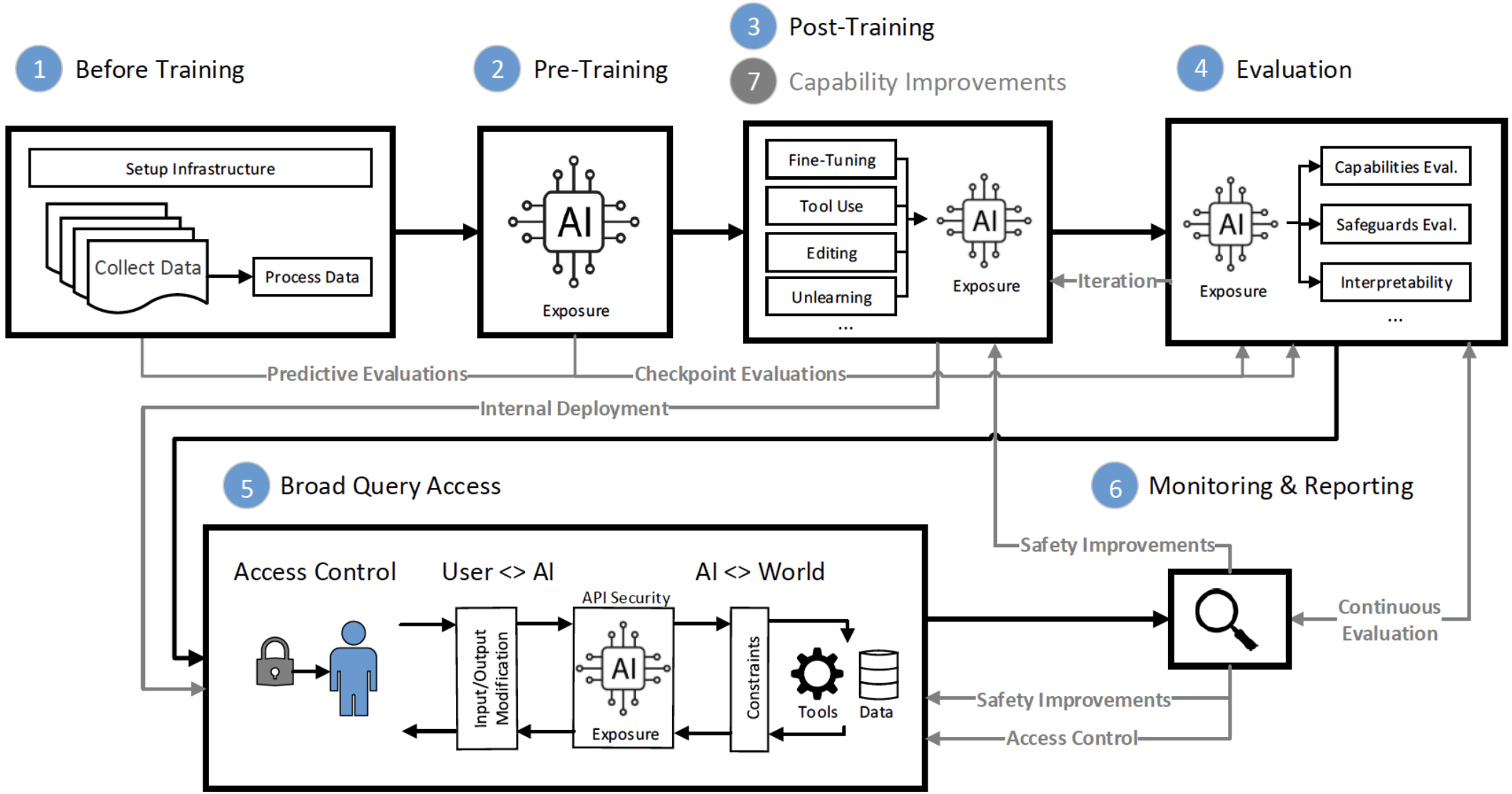

Model Poisoning
Model Control
Energy Latency
Query Access
Availability
Label Limit
Clean-Label Poisoning
Training Data Control
Data Poisoning

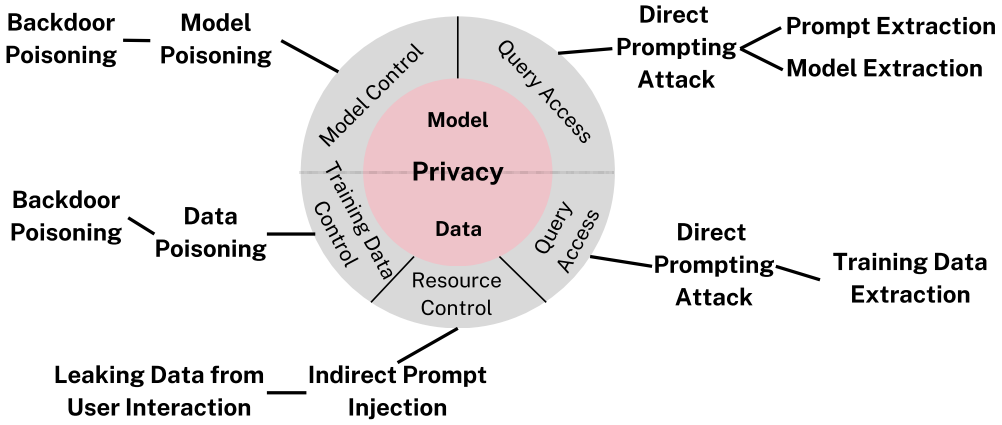

Clean-Label Poisoning
Clean-Label Backdoor
Model Poisoning
Label Limit
Model Control
Source Code Control
Integrity
Query Access
Black-Box Evasion
Backdoor Poisoning
Test Data Control
Training Data Control
Backdoor Poisoning
Evasion
Targeted Poisoning

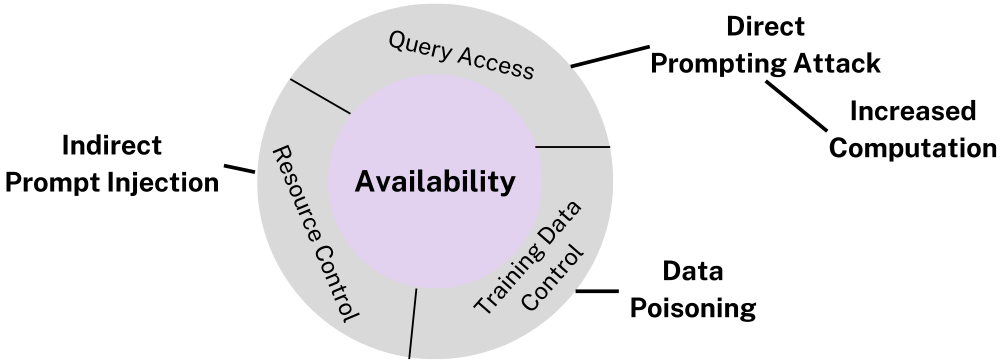

Model Extraction
Query Access
Model
Privacy
Data
Query Access
Reconstruction; Membership Inference; Property Inference

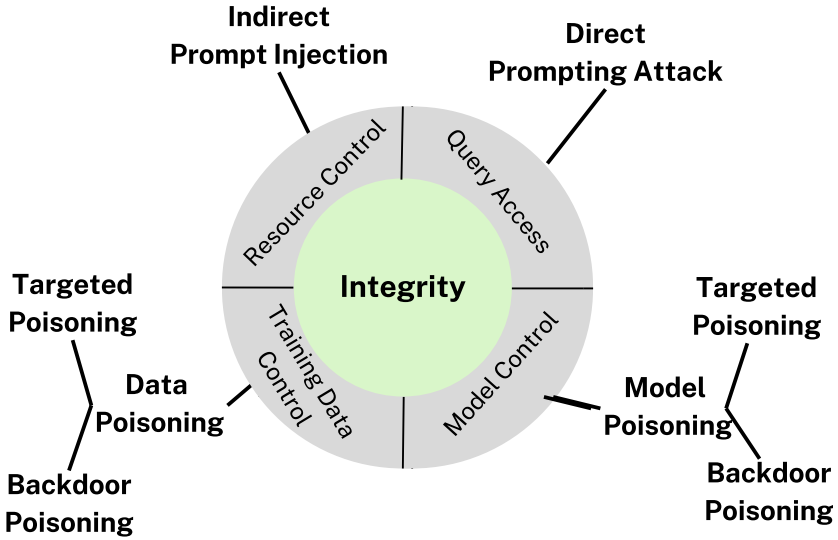

Model Extraction
Query Access
Model
Privacy
Data
Query Access
Reconstruction;
Membership Inference;
Property Inference

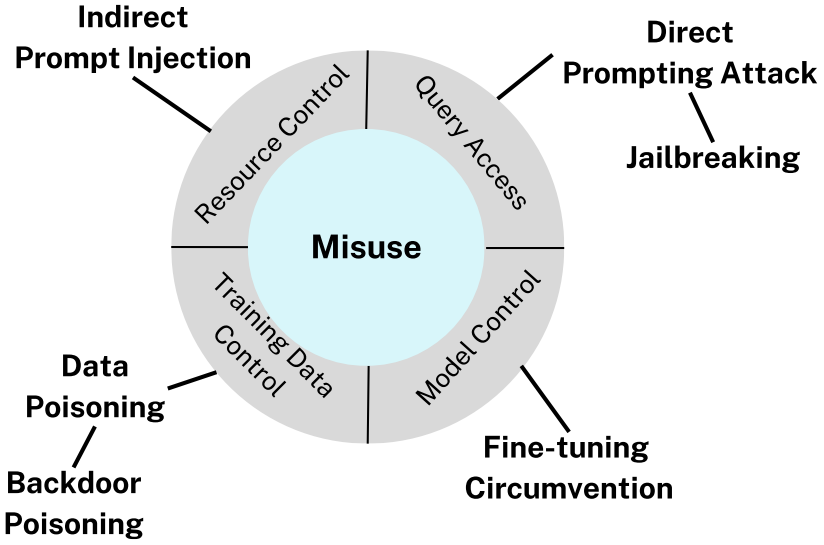
GenAI

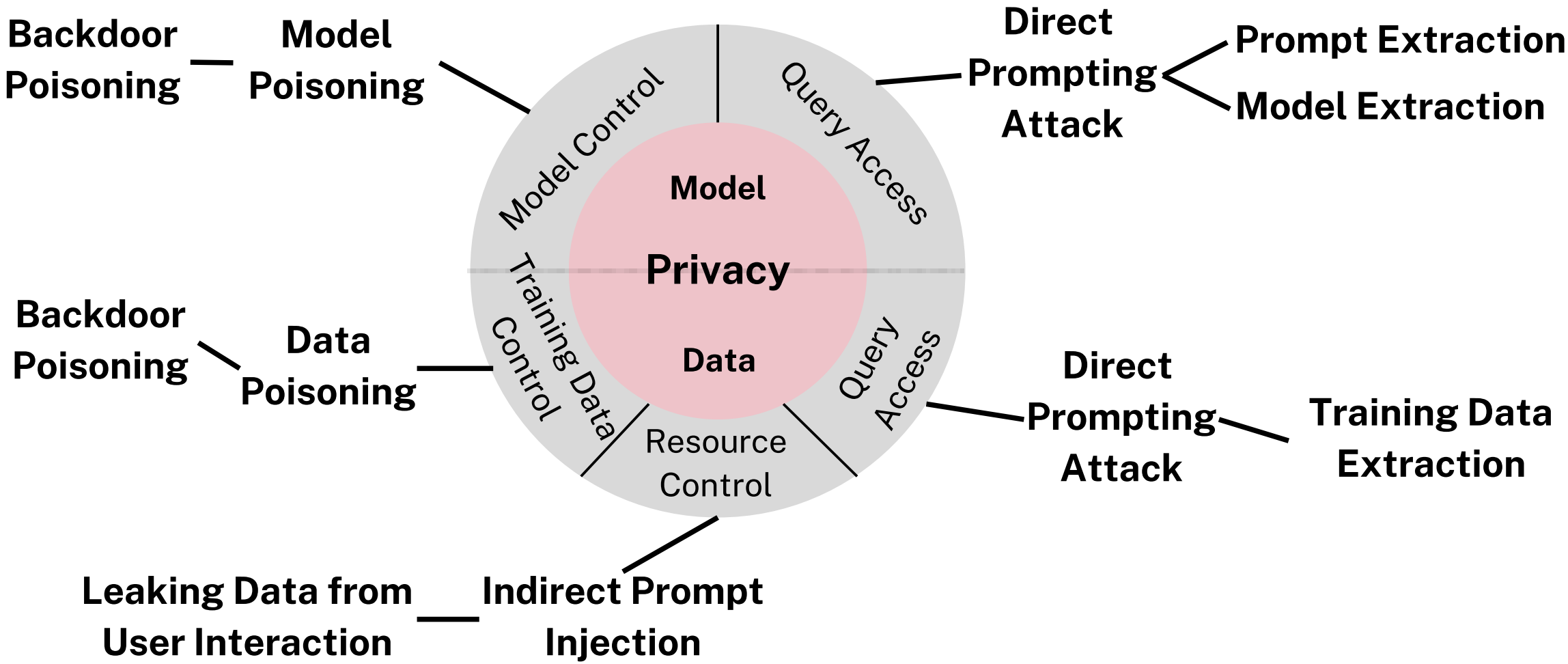

1 Before Training
Setup Infrastructure
Collect Data
Process Data
2 Pre-Training
AI
Exposure
3 Post-Training
7 Capability Improvements
Fine-Tuning
Tool Use
Editing
Unlearning
...
AI
Exposure
4 Evaluation
AI
Exposure
Capabilities Eval.
Safeguards Eval.
Interpretability
...
Iteration
Predictive Evaluations
Checkpoint Evaluations
Internal Deployment
5 Broad Query Access
Access Control
User <> AI
AI <> World
API Security
Input/Output Modification
AI
Exposure
Constraints
Tools
Data
6 Monitoring & Reporting
Safety Improvements
Continuous Evaluation
Safety Improvements
Access Control

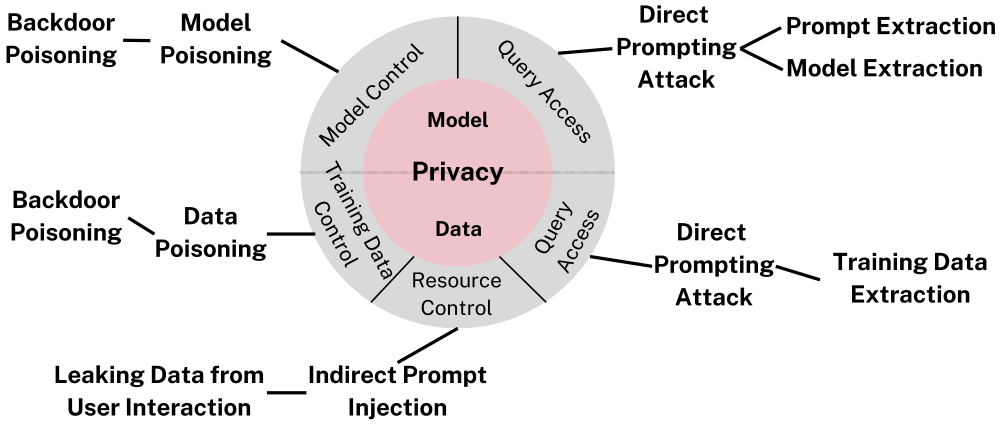

Backdoor Poisoning
Model Poisoning
Model Control
Query Access
Direct Prompting Attack
Prompt Extraction
Model Extraction
Model
Privacy
Data
Backdoor Poisoning
Data Poisoning
Training Data Control
Query Access
Direct Prompting Attack
Training Data Extraction
Resource Control
Leaking Data from User Interaction
Indirect Prompt Injection

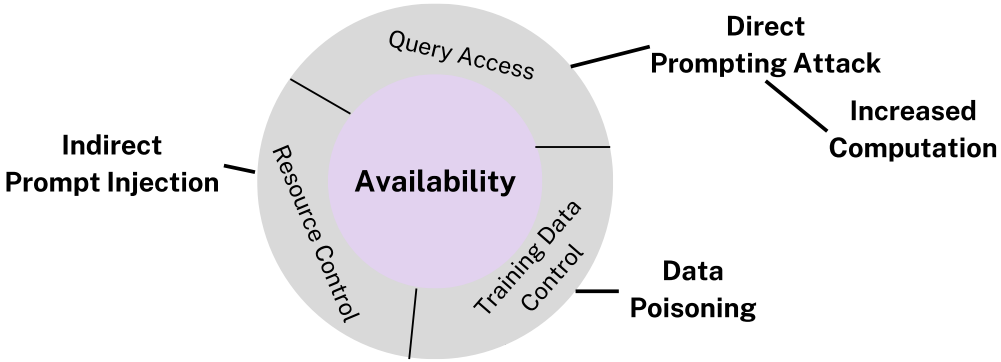

Query Access
Direct Prompting Attack
Increased Computation
Indirect Prompt Injection
Resource Control
Availability
Training Data Control
Data Poisoning

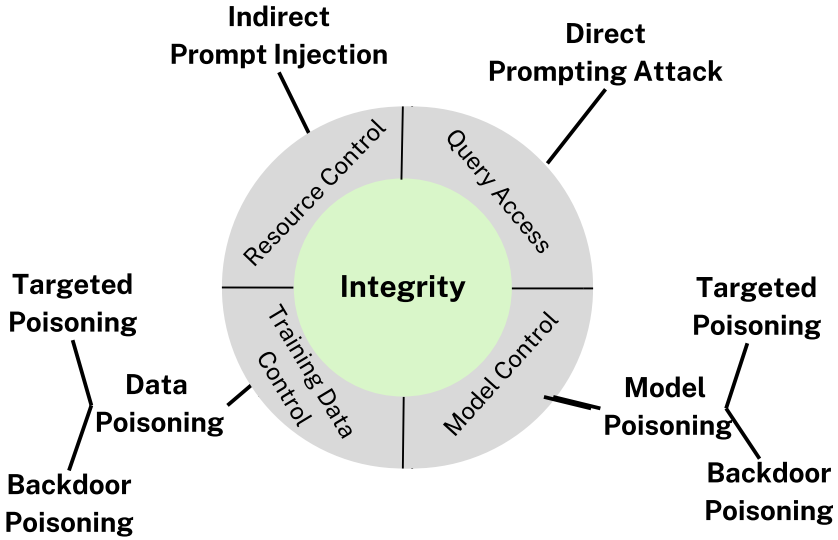

Indirect Prompt Injection
Direct Prompting Attack
Resource Control
Query Access
Integrity
Targeted Poisoning
Training Data Control
Model Control
Targeted Poisoning
Data Poisoning
Model Poisoning
Backdoor Poisoning
Backdoor Poisoning

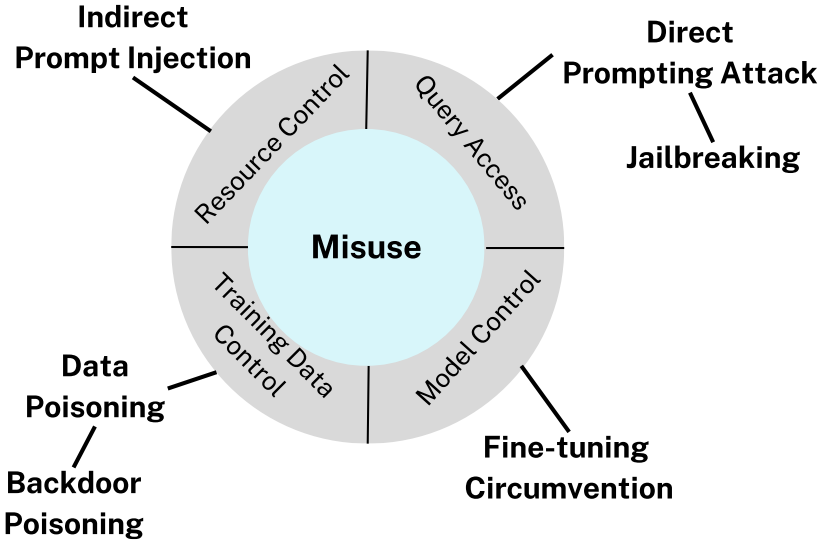

Indirect Prompt Injection
Direct Prompting Attack
Jailbreaking
Resource Control
Query Access
Misuse
Training Data Control
Model Control
Data Poisoning
Backdoor Poisoning
Fine-tuning Circumvention

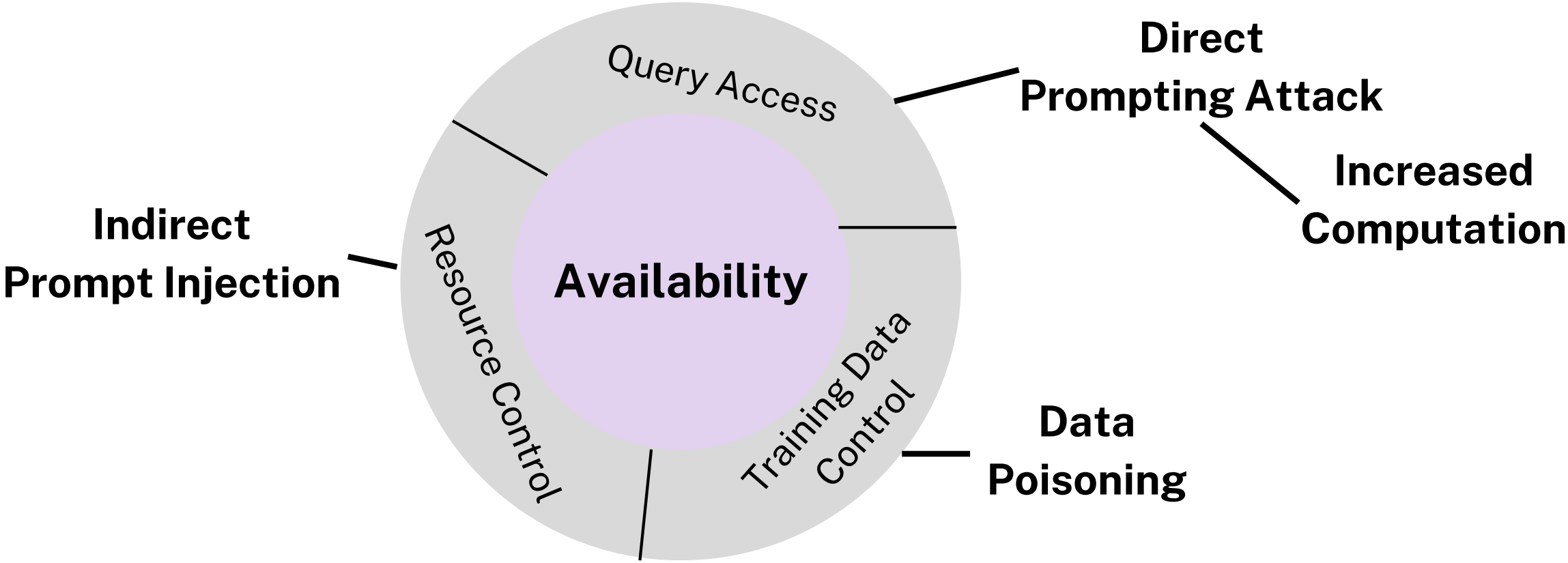

Backdoor Poisoning
Model Poisoning
Direct Prompting Attack
Prompt Extraction
Model Extraction
Model Control
Query Access
Model
Privacy
Data
Training Data Control
Query Access
Resource Control
Backdoor Poisoning
Data Poisoning
Direct Prompting Attack
Training Data Extraction
Leaking Data from User Interaction
Indirect Prompt Injection

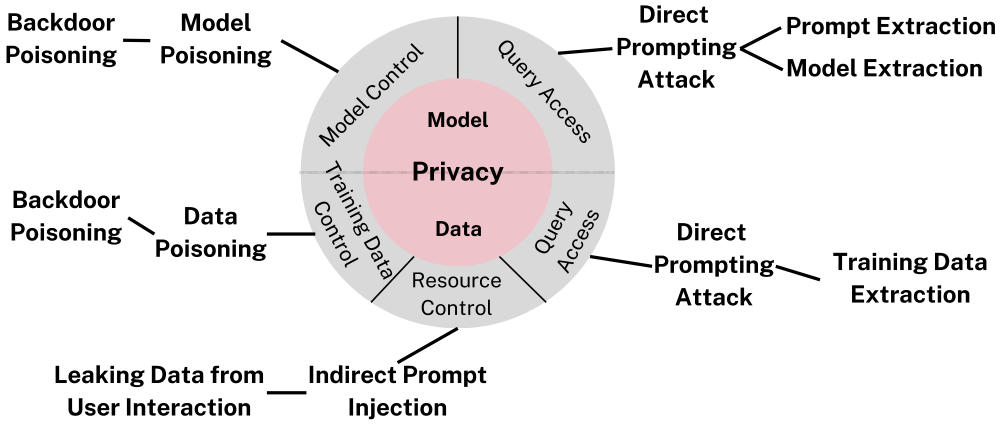

Backdoor Poisoning
Model Poisoning
Model Control
Query Access
Direct Prompting Attack
Prompt Extraction
Model Extraction
Model
Privacy
Data
Backdoor Poisoning
Data Poisoning
Training Data Control
Query Access
Direct Prompting Attack
Training Data Extraction
Resource Control
Leaking Data from User Interaction
Indirect Prompt Injection

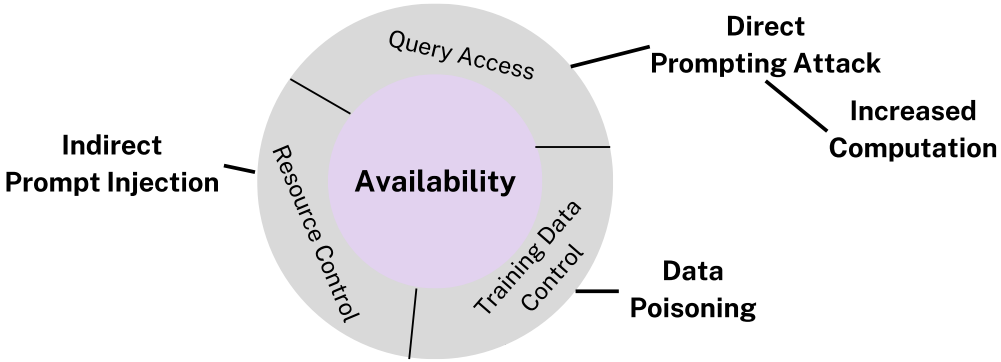

Query Access
Direct Prompting Attack
Increased Computation
Indirect Prompt Injection
Resource Control
Availability
Training Data Control
Data Poisoning

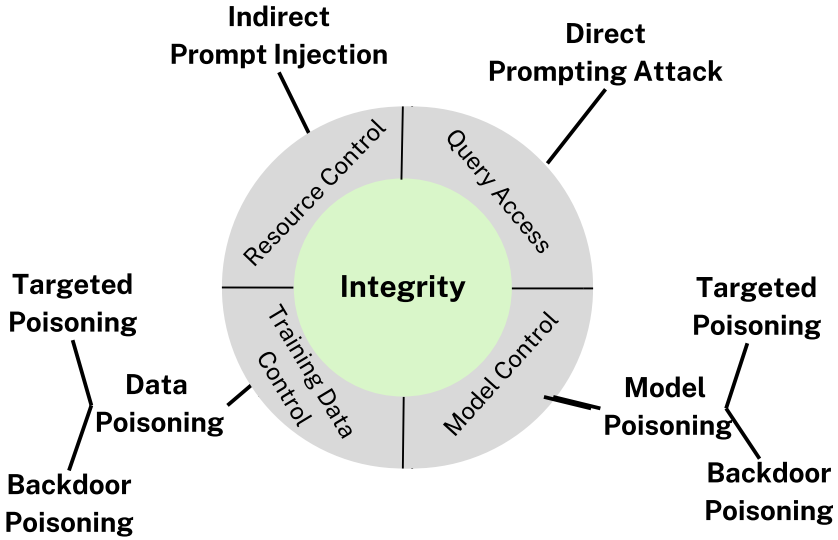

Indirect Prompt Injection
Direct Prompting Attack
Resource Control
Query Access
Integrity
Targeted Poisoning
Training Data Control
Model Control
Targeted Poisoning
Data Poisoning
Model Poisoning
Backdoor Poisoning
Backdoor Poisoning

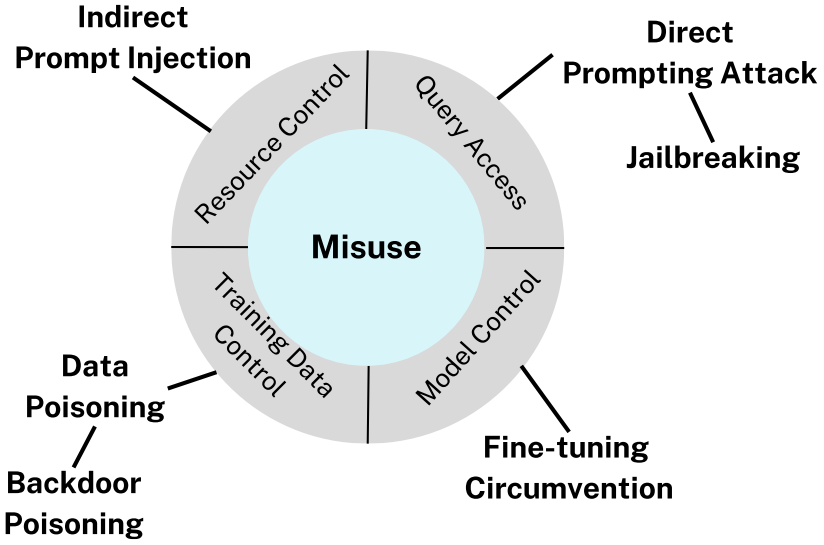

Indirect Prompt Injection
Direct Prompting Attack
Jailbreaking
Resource Control
Query Access
Misuse
Training Data Control
Model Control
Data Poisoning
Backdoor Poisoning
Fine-tuning Circumvention

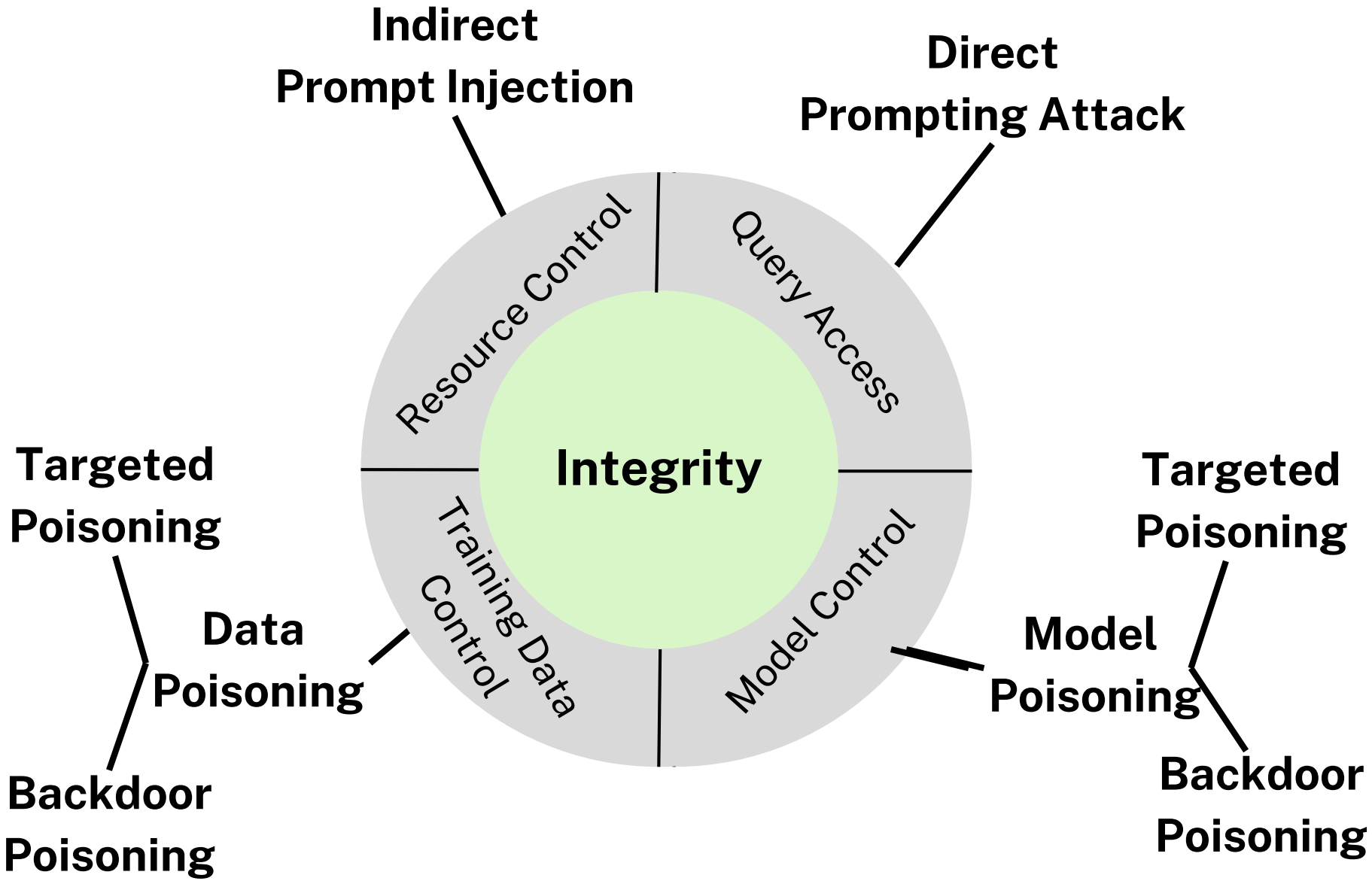

Query Access
Resource Control
Training Data Control
Availability
Direct Prompting Attack
Increased Computation
Indirect Prompt Injection
Data Poisoning

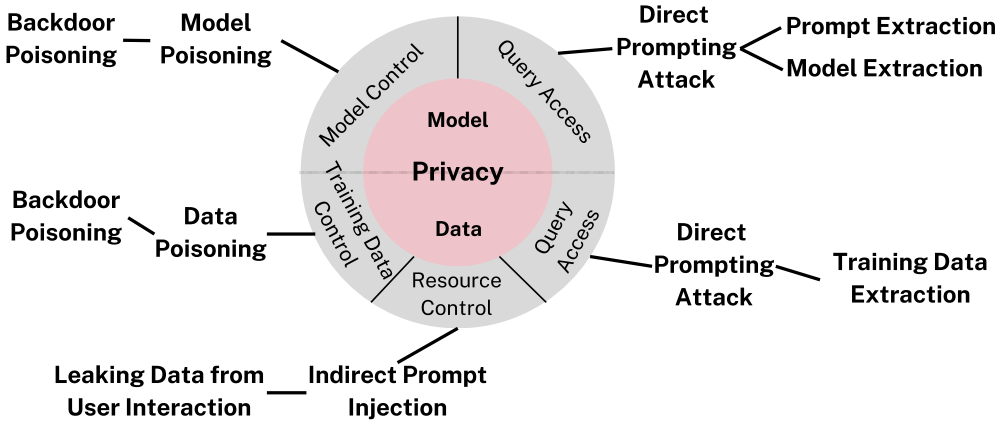

Backdoor Poisoning
Model Poisoning
Model Control
Query Access
Direct Prompting Attack
Prompt Extraction
Model Extraction
Model
Privacy
Data
Backdoor Poisoning
Data Poisoning
Training Data Control
Query Access
Direct Prompting Attack
Training Data Extraction
Resource Control
Leaking Data from User Interaction
Indirect Prompt Injection

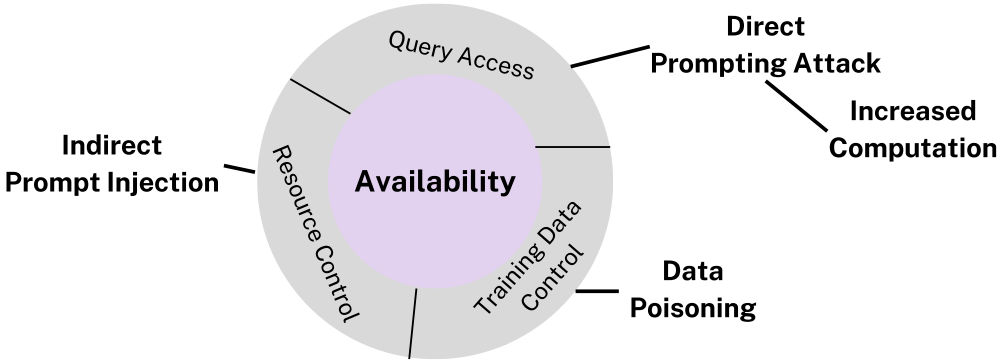

Query Access
Direct Prompting Attack
Increased Computation
Indirect Prompt Injection
Resource Control
Availability
Training Data Control
Data Poisoning

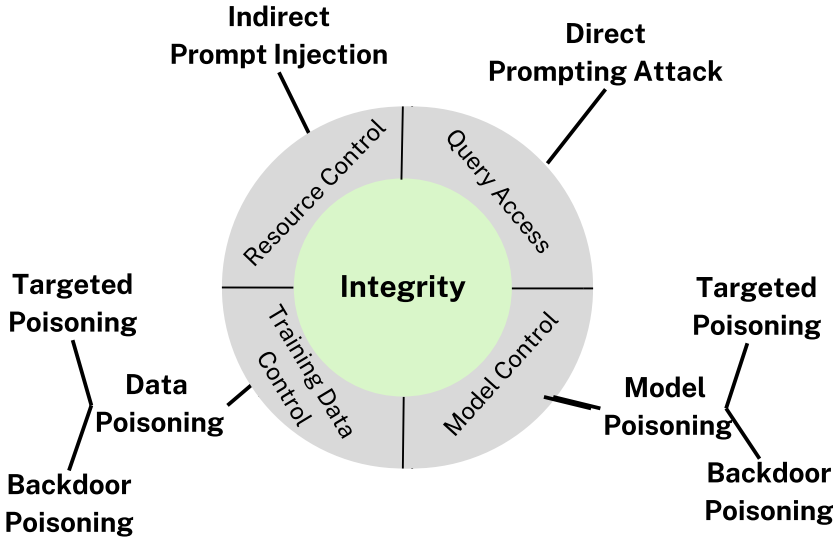

Indirect Prompt Injection
Direct Prompting Attack
Resource Control
Query Access
Integrity
Targeted Poisoning
Training Data Control
Model Control
Targeted Poisoning
Data Poisoning
Model Poisoning
Backdoor Poisoning
Backdoor Poisoning

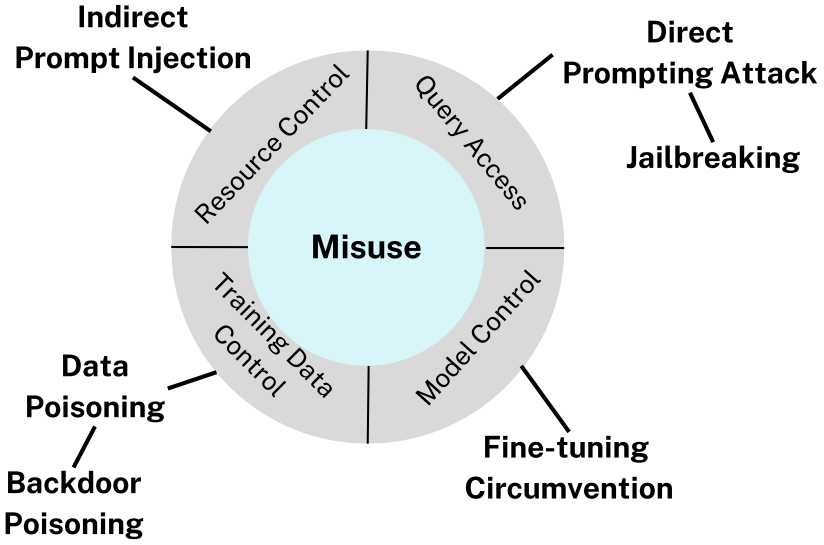

Indirect Prompt Injection
Direct Prompting Attack
Jailbreaking
Resource Control
Query Access
Misuse
Training Data Control
Model Control
Data Poisoning
Backdoor Poisoning
Fine-tuning Circumvention

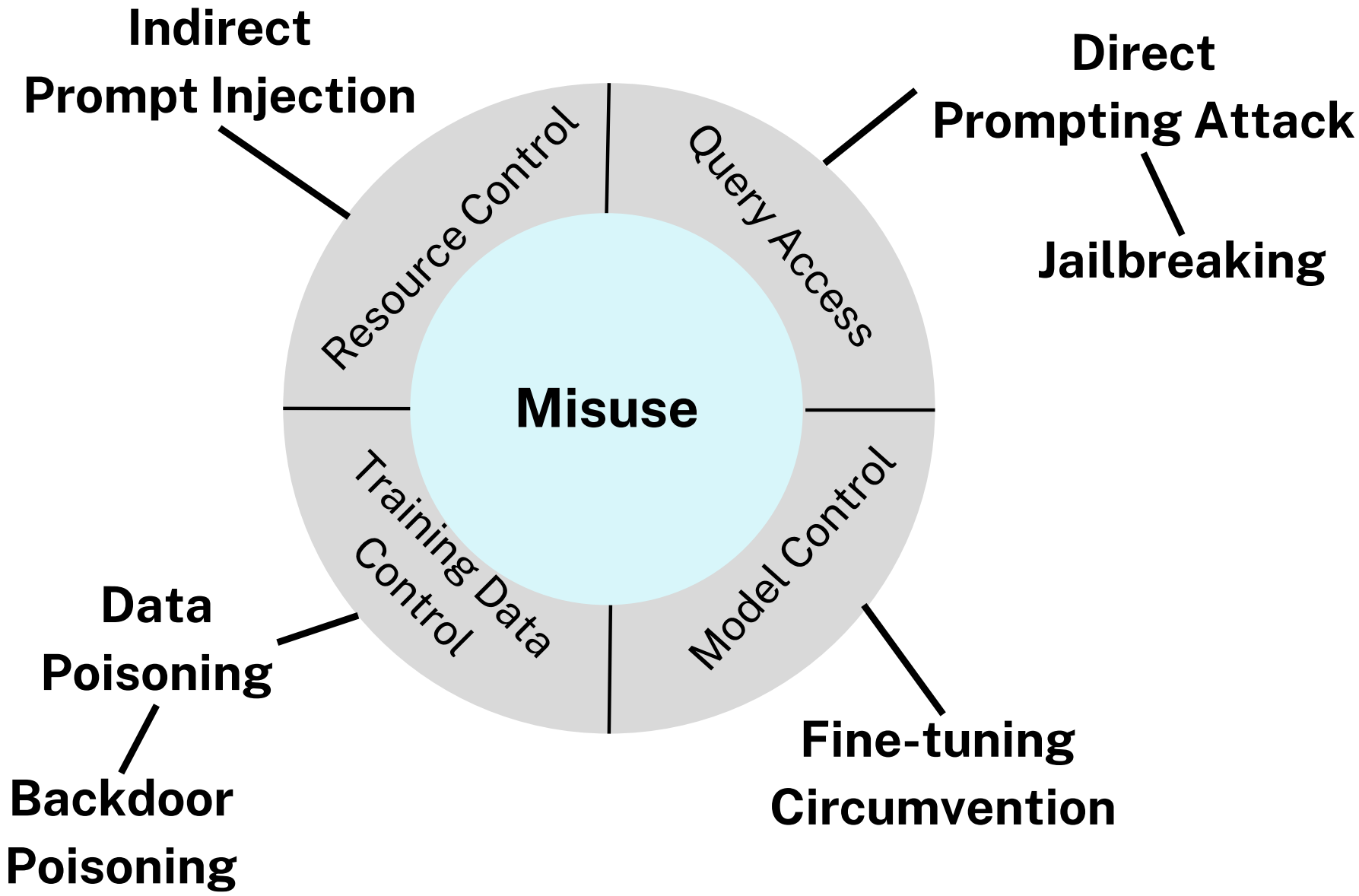

Indirect
Prompt Injection
Direct
Prompting Attack
Resource Control
Query Access
Integrity
Training Data
Control
Model Control
Targeted
Poisoning
Data
Poisoning
Backdoor
Poisoning
Model
Poisoning
Targeted
Poisoning
Backdoor
Poisoning

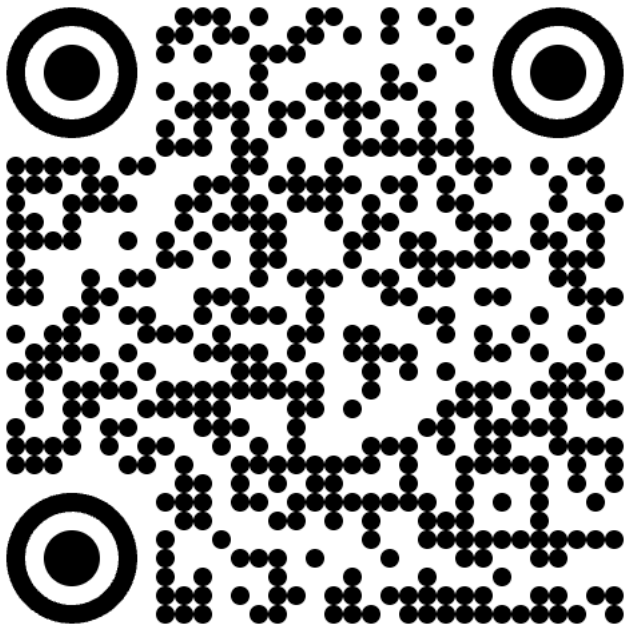

Backdoor Poisoning
Model Poisoning
Model Control
Query Access
Direct Prompting Attack
Prompt Extraction
Model Extraction
Model
Privacy
Data
Backdoor Poisoning
Data Poisoning
Training Data Control
Query Access
Direct Prompting Attack
Training Data Extraction
Resource Control
Leaking Data from User Interaction
Indirect Prompt Injection


Query Access
Direct Prompting Attack
Increased Computation
Indirect Prompt Injection
Resource Control
Availability
Training Data Control
Data Poisoning


Indirect Prompt Injection
Direct Prompting Attack
Resource Control
Query Access
Integrity
Targeted Poisoning
Training Data Control
Model Control
Targeted Poisoning
Data Poisoning
Model Poisoning
Backdoor Poisoning
Backdoor Poisoning


Indirect Prompt Injection
Direct Prompting Attack
Jailbreaking
Resource Control
Query Access
Misuse
Training Data Control
Model Control
Data Poisoning
Backdoor Poisoning
Fine-tuning Circumvention

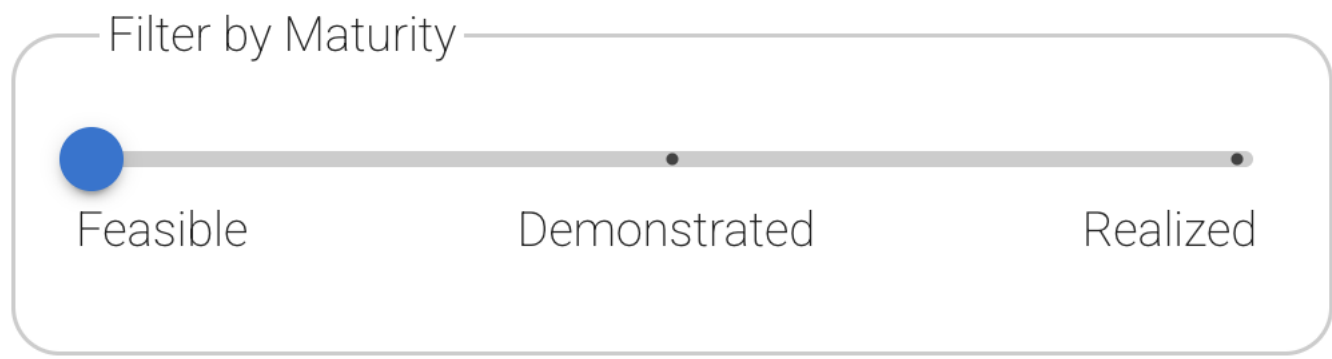

Indirect
Prompt Injection
Direct
Prompting Attack
Jailbreaking
Resource Control
Query Access
Misuse
Training Data
Control
Model Control
Data
Poisoning
Backdoor
Poisoning
Fine-tuning
Circumvention



https://atlas.mitre.org/matrices/ATLAS

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML tec
about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View th



Reconnaissance[&]	Resource Development[&]	Initial Access[&]	AI Model Access	Execution[&]	Persistence[&]	Privilege Escalation[&]	Defense Evasion[&]	Credential Access[&]	Discovery[&]	Lateral Movement[&]	Collection[&]	AI Attack Staging	Command and Control[&]	Exfiltration[&]	Impact[&]
8 techniques	13 techniques	7 techniques	4 techniques	6 techniques	8 techniques	4 techniques	13 techniques	6 techniques	9 techniques	2 techniques	4 techniques	6 techniques	3 techniques	6 techniques	8 techniques
Active Scanning [&]	Acquire Infrastructure	AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Clickbait	AI Agent Context Poisoning	AI Agent Tool Invocation	Corrupt AI Model	AI Agent Tool Credential Harvesting	Cloud Service Discovery [&]	Phishing [&]	AI Artifact Collection	Craft Adversarial Data	AI Agent	Exfiltration via AI Agent Tool Invocation	Cost Harvesting
Gather RAG-Indexed Targets	Acquire Public AI Artifacts	Drive-by Compromise [&]	AI-Enabled Product or Service	AI Agent Tool Invocation	AI Agent Tool Data Poisoning	Escape to Host [&]	Delay Execution of LLM Instructions	Credentials from AI Agent Configuration	Discover AI Agent Configuration	Use Alternate Authentication Material [&]	Data from AI Services	Create Proxy AI Model	AI Service API	Exfiltration via AI Inference API	Data Destruction via AI Agent Tool Invocation
Gather Victim Identity Information [&]	Develop Capabilities [&]	Evade AI Model	Full AI Model Access	Command and Scripting Interpreter [&]	LLM Prompt Self-Replication	LLM Jailbreak	Evade AI Model	Exploitation for Credential Access [&]	Discover AI Artifacts		Data from Information Repositories [&]	Generate Deepfakes	Reverse Shell	Exfiltration via Cyber Means	Denial of AI Service
Search Application Repositories	Establish Accounts [&]	Exploit Public-Facing Application [&]	Physical Environment Access	Deploy AI Agent	Manipulate AI Model	Valid Accounts [&]	Exploitation for Defense Evasion [&]	OS Credential Dumping [&]	Discover AI Model Family		Data from Local System [&]	Generate Malicious Commands		Extract LLM System Prompt	Erode AI Model Integrity
Search Open AI Vulnerability Analysis	LLM Prompt Crafting	Phishing [&]		LLM Prompt Injection	Modify AI Agent Configuration		False RAG Entry Injection	RAG Credential Harvesting	Discover AI Model Ontology			Manipulate AI Model		LLM Data Leakage	Erode Dataset Integrity
Search Open Technical Databases [&]	Obtain Capabilities [&]	Prompt Infiltration via Public-Facing Application		User Execution [&]	Poison Training Data		Impersonation [&]	Unsecured Credentials [&]	Discover AI Model Outputs			Verify Attack		LLM Response Rendering	Evade AI Model
Search Open Websites/Domains [&]	Poison Training Data	Valid Accounts [&]			Prompt Infiltration via Public-Facing Application		LLM Jailbreak		Discover LLM Hallucinations						External Harms
Search Victim-Owned Websites [&]	Publish Hallucinated Entities				RAG Poisoning		LLM Prompt Obfuscation		Discover LLM System Information						Spamming AI System with Chaff Data
	Publish Poisoned AI Agent Tool						LLM Trusted Output Components Manipulation		Process Discovery [&]						
	Publish Poisoned Datasets						Manipulate User LLM Chat History								
	Publish Poisoned Models						Masquerading [&]								
	Retrieval Content Crafting						Modify AI Agent Configuration								
	Stage Capabilities [&]						Virtualization/Sandbox Evasion [&]								

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

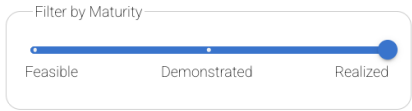
Filter by Maturity

Feasible — Demonstrated — Realized

Reconnaissance &	Resource Development &	Initial Access &	AI Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Lateral Movement &	Collection &	AI Attack Staging	Command and Control &	Exfiltration &	Impact &
8 techniques	13 techniques	7 techniques	4 techniques	5 techniques	7 techniques	4 techniques	13 techniques	6 techniques	8 techniques	2 techniques	4 techniques	6 techniques	3 techniques	4 techniques	6 techniques
Active Scanning &	Acquire Infrastructure	AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Tool Invocation	AI Agent Context Poisoning	AI Agent Tool Invocation	Corrupt AI Model	AI Agent Tool Credential Harvesting	Cloud Service Discovery &	Phishing &	AI Artifact Collection	Craft Adversarial Data	AI Agent	Exfiltration via AI Agent Tool Invocation	Data Destruction via AI Agent Tool Invocation
Gather RAG-Indexed Targets	Acquire Public AI Artifacts	Drive-by Compromise &	AI-Enabled Product or Service	Command and Scripting Interpreter &	LLM Prompt Self-Replication	Escape to Host &	Delay Execution of LLM Instructions	Credentials from AI Agent Configuration	Discover AI Agent Configuration	Use Alternate Authentication Material &	Data from AI Services	Create Proxy AI Model	AI Service API	Exfiltration via Cyber Means	Denial of AI Service
Gather Victim Identity Information &	Develop Capabilities &	Evade AI Model	Full AI Model Access	Deploy AI Agent	Manipulate AI Model	LLM Jailbreak	Evade AI Model	Exploitation for Credential Access &	Discover AI Artifacts		Data from Information Repositories &	Generate Deepfakes	Reverse Shell	LLM Data Leakage	Erode AI Model Integrity
Search Application Repositories	Establish Accounts &	Exploit Public-Facing Application &	Physical Environment Access	LLM Prompt Injection	Modify AI Agent Configuration	Valid Accounts &	Exploitation for Defense Evasion &	OS Credential Dumping &	Discover AI Model Ontology		Data from Local System &	Generate Malicious Commands		LLM Response Rendering	Erode Dataset Integrity
Search Open AI Vulnerability Analysis	LLM Prompt Crafting	Phishing &		User Execution &	Poison Training Data		False RAG Entry Injection	RAG Credential Harvesting	Discover AI Model Outputs			Manipulate AI Model			Evade AI Model
Search Open Technical Databases &	Obtain Capabilities &	Prompt Infiltration via Public-Facing Application			Prompt Infiltration via Public-Facing Application		Impersonation &	Unsecured Credentials &	Discover LLM Hallucinations			Verify Attack			External Harms
Search Open Websites/Domains &	Poison Training Data	Valid Accounts &			RAG Poisoning		LLM Jailbreak		Discover LLM System Information						
Search Victim-Owned Websites &	Publish Hallucinated Entities						LLM Prompt Obfuscation		Process Discovery &						
	Publish Poisoned AI Agent Tool						LLM Trusted Output Components Manipulation								
	Publish Poisoned Datasets						Manipulate User LLM Chat History								
	Publish Poisoned Models						Masquerading &								
	Retrieval Content Crafting						Modify AI Agent Configuration								
	Stage Capabilities &						Virtualization/Sandbox Evasion &								

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.



Reconnaissance &	Resource Development &	Initial Access &	AI Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Lateral Movement &	Collection &	AI Attack Staging	Command and Control &	Exfiltration &	Impact &
2 techniques	8 techniques	5 techniques	1 technique	3 techniques	2 techniques	1 technique	5 techniques	1 technique	1 technique	1 technique	3 techniques	4 techniques	2 techniques	1 technique	4 techniques
Active Scanning &	Acquire Infrastructure	AI Supply Chain Compromise	AI-Enabled Product or Service	Deploy AI Agent	Manipulate AI Model	Valid Accounts &	Corrupt AI Model	Unsecured Credentials &	Cloud Service Discovery &	Phishing &	AI Artifact Collection	Craft Adversarial Data	AI Service API	Exfiltration via Cyber Means	Data Destruction via AI Agent Tool Invocation
Gather Victim Identity Information &	Acquire Public AI Artifacts	Evade AI Model		LLM Prompt Injection	Poison Training Data		Evade AI Model				Data from Information Repositories &	Generate Deepfakes	Reverse Shell		Erode AI Model Integrity
	Develop Capabilities &	Exploit Public-Facing Application &		User Execution &			Impersonation &				Data from Local System &	Generate Malicious Commands			Evade AI Model
	Establish Accounts &	Phishing &					Masquerading &					Manipulate AI Model			External Harms
	LLM Prompt Crafting	Valid Accounts &					Virtualization/Sandbox Evasion &								
	Obtain Capabilities &														
	Poison Training Data														
	Publish Poisoned Models														

LLM Prompt Injection

An adversary may craft malicious prompts as inputs to an LLM that cause the LLM to act in unintended ways. These "prompt injections" are often designed to cause the model to ignore aspects of its original instructions and follow the adversary's instructions instead.

Prompt Injections can be an initial access vector to the LLM that provides the adversary with a foothold to carry out other steps in their operation. They may be designed to bypass defenses in the LLM, or allow the adversary to issue privileged commands. The effects of a prompt injection can persist throughout an interactive session with an LLM.

Malicious prompts may be injected directly by the adversary (Direct) either to leverage the LLM to generate harmful content or to gain a foothold on the system and lead to further effects. Prompts may also be injected indirectly when as part of its normal operation the LLM ingests the malicious prompt from another data source (Indirect). This type of injection can be used by the adversary to a foothold on the system or to target the user of the LLM. Malicious prompts may also be Triggered user actions or system events.

ID: AML.T0051

Subtechniques: LLM Prompt Injection: Direct, LLM Prompt Injection: Indirect, LLM Prompt Injection: Triggered

Tactic: Execution

Maturity: Realized

Number Of Case Studies: 1

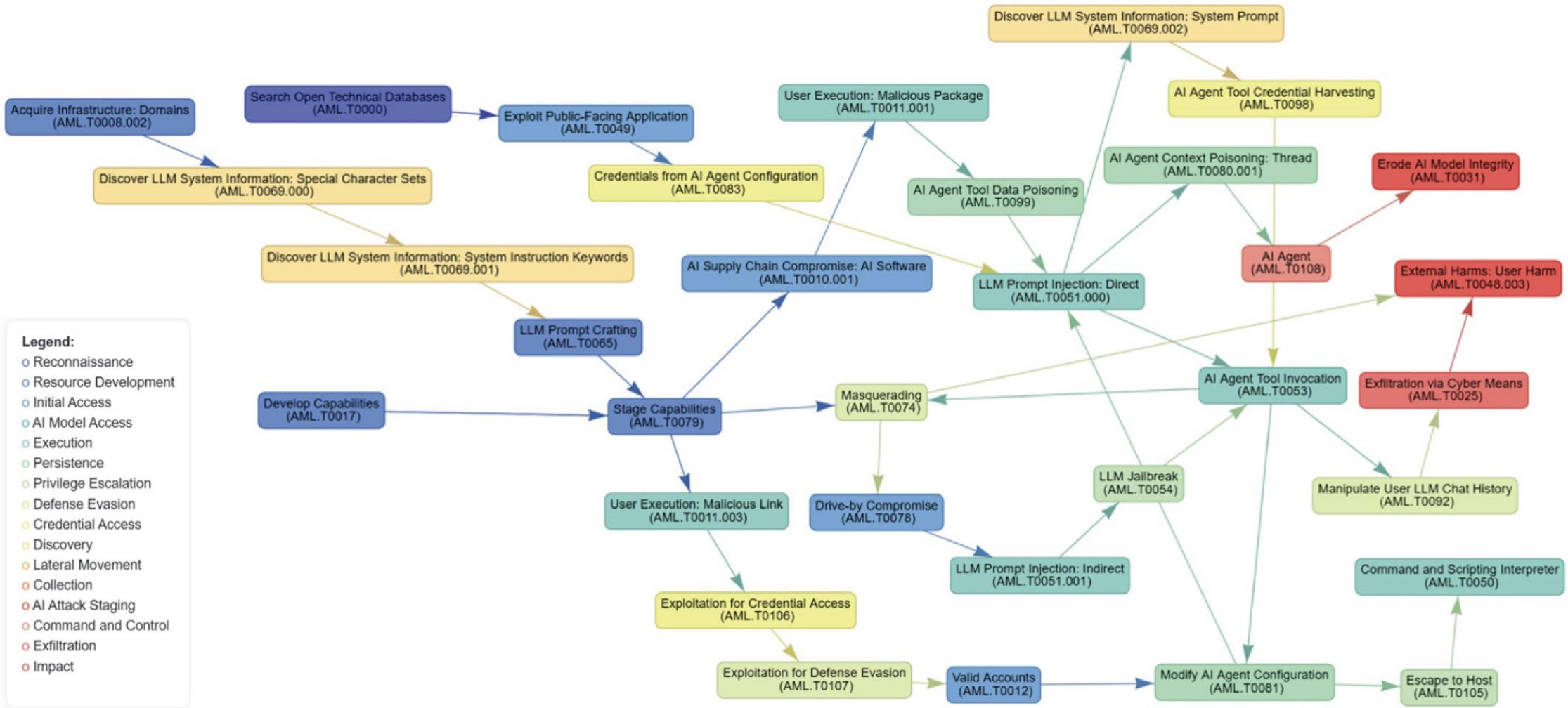
Number Of Mitigations: 6

Created: 25 October 2023

Last Modified: 05 November 2025

MITRE ATLAS
OPENCLAW INVESTIGATION

https://www.mitre.org/news-insights/publication/mitre-atlas-openclaw-investigation



OPENCLAW ATTACK GRAPH

Microsoft

https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf

Taxonomy of Failure Mode in Agentic AI Systems

	Safety	Security
Novel	• Intra-agent Responsible AI (RAI) issues • Harms of allocation in multi-user scenarios • Organizational knowledge loss • Prioritization leading to user safety issues	• Agent compromise • Agent injection • Agent impersonation • Agent flow manipulation • Agent provisioning poisoning • Multi-agent jailbreaks
Existing	• Insufficient transparency and accountability • Parasocial relationships • Bias amplification • User impersonation • Insufficient intelligibility for meaningful consent • Hallucinations • Misinterpretation of instructions	• Memory poisoning and theft • Targeted knowledge base poisoning • XPIA • Human-in-the-loop bypass • Function compromise and malicious functions • Incorrect permissions • Resource exhaustion • Insufficient isolation • Excessive agency • Loss of data provenance

Image: Microsoft Designer

Build | Test | Deploy | Operate

Modelscan
Scans downloaded model weights.
https://github.com/protectai/modelscan

ZenML
Tracking lineage of data and models.
https://github.com/zenml-io/zenml

Image:
Microsoft Designer

Nvidia NeMo Guardrails
Rules for restricting interactions.
https://github.com/NVIDIA-NeMo/Guardrails

META Llama Guard
Specialised LLM as security classifier.
https://huggingface.co/meta-llama/Llama-Guard-4-12B

Langfuse
Visual traces of LLM and agent workflows.
https://github.com/langfuse/langfuse

Arize Phoenix
Context poisoning monitoring/detector.
https://phoenix.arize.com/

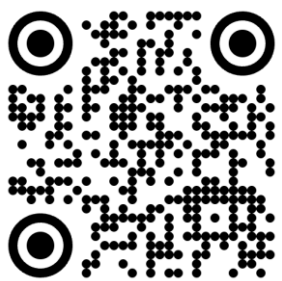
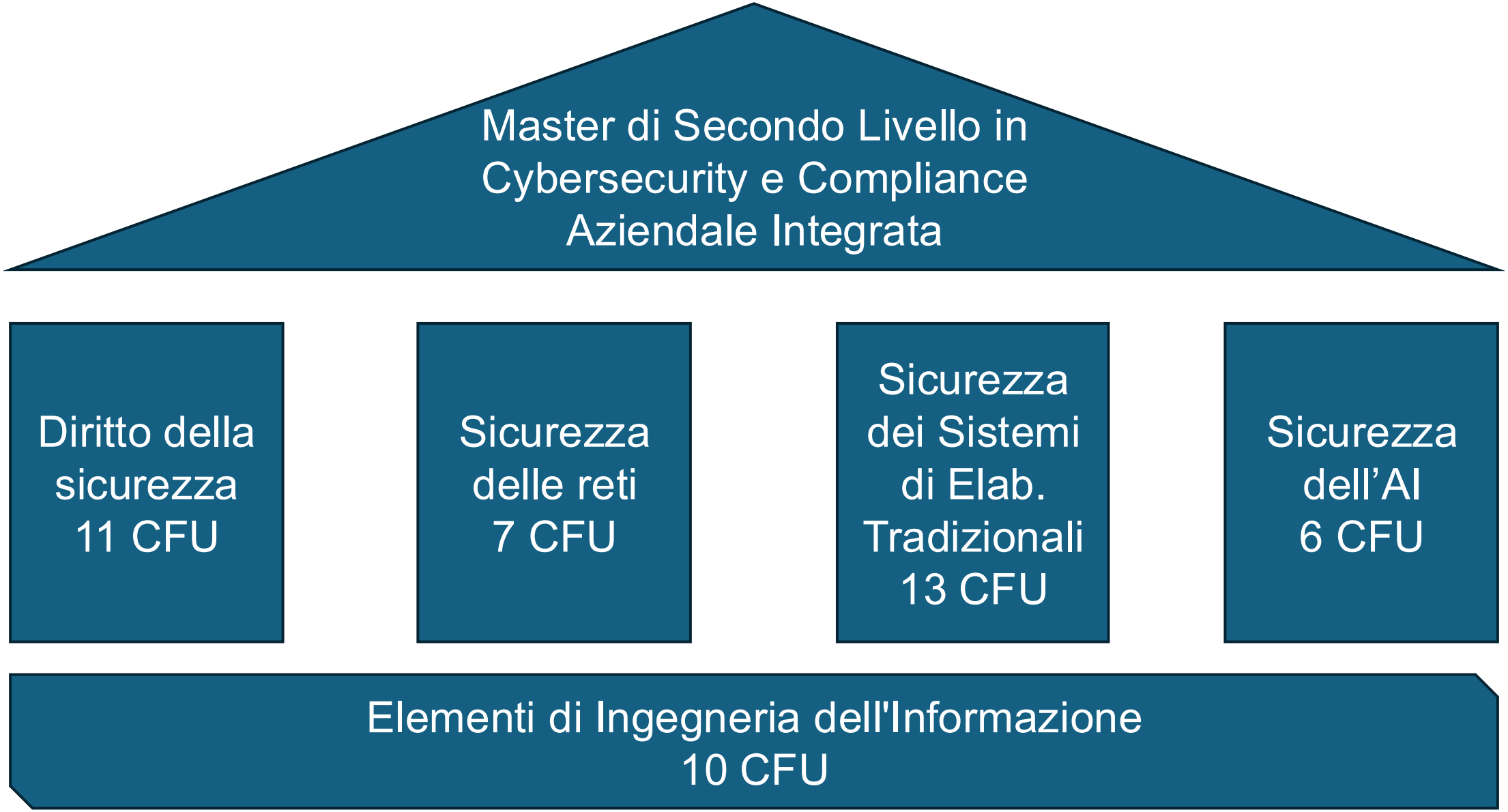
NVIDIA Garak	Microsoft PyRIT	Promptfoo	DeepTeam
The automated vulnerability scanner (Nmap for LLMs).	The advanced persistent threat simulator.	The DevSecOps pipeline defender.	The out-of-the-box compliance checker.
Fast, broad-strokes probing for known CVEs and basic jailbreaks.	Multi-turn, compounding conversational attacks and multi-modal testing.	Lightweight, YAML-based testing that runs automatically on every code push.	Pre-loaded with 40+ vulnerabilities; automatically maps to OWASP and NIST.
Security Auditors establishing initial safety baselines.	Dedicated Red Teams testing complex AI agents.	Developers & QA catching regressions during daily builds.	Governance & Risk (GRC) teams needing quick compliance reports.
https://garak.ai/	https://github.com/Azure/PyRIT	https://promptfoo.dev/	https://trydeepteam.com/
			

Image: ChatGPT 5.2


Master di Secondo Livello in
Cybersecurity e Compliance
Aziendale Integrata
Diritto della
sicurezza
11 CFU
Sicurezza
delle reti
7 CFU
Sicurezza
dei Sistemi
di Elab.
Tradizionali
13 CFU
Sicurezza
dell'AI
6 CFU
Elementi di Ingegneria dell'Informazione
10 CFU

Federico Cerutti
federico.cerutti@unibs.it