Known Attacks and Their Impacts on AI Systems

2nd Edition April 2026

1st Edition March 2025

AISI Japan AI Safety Institute

Abstract

- This document provides an overview of attacks specific to AI systems and their impacts.
- As AI systems evolve, **attacks exploiting their characteristics** are increasingly reported, leading to issues like **training data leaks** and **abnormal AI outputs**.
- In the real world, these issues affect people's lives:
  - **Leaked training data** from AI used in medical diagnosis violates privacy.
  - **Abnormal AI outputs** in automated driving can cause traffic accidents.
- Thus, AI-specific security measures are essential.
- This document outlines attacks and their impacts, as reported in academic papers and other sources, to aid in examining countermeasures.

Scope of This Document

AISI Japan AI Safety Institute

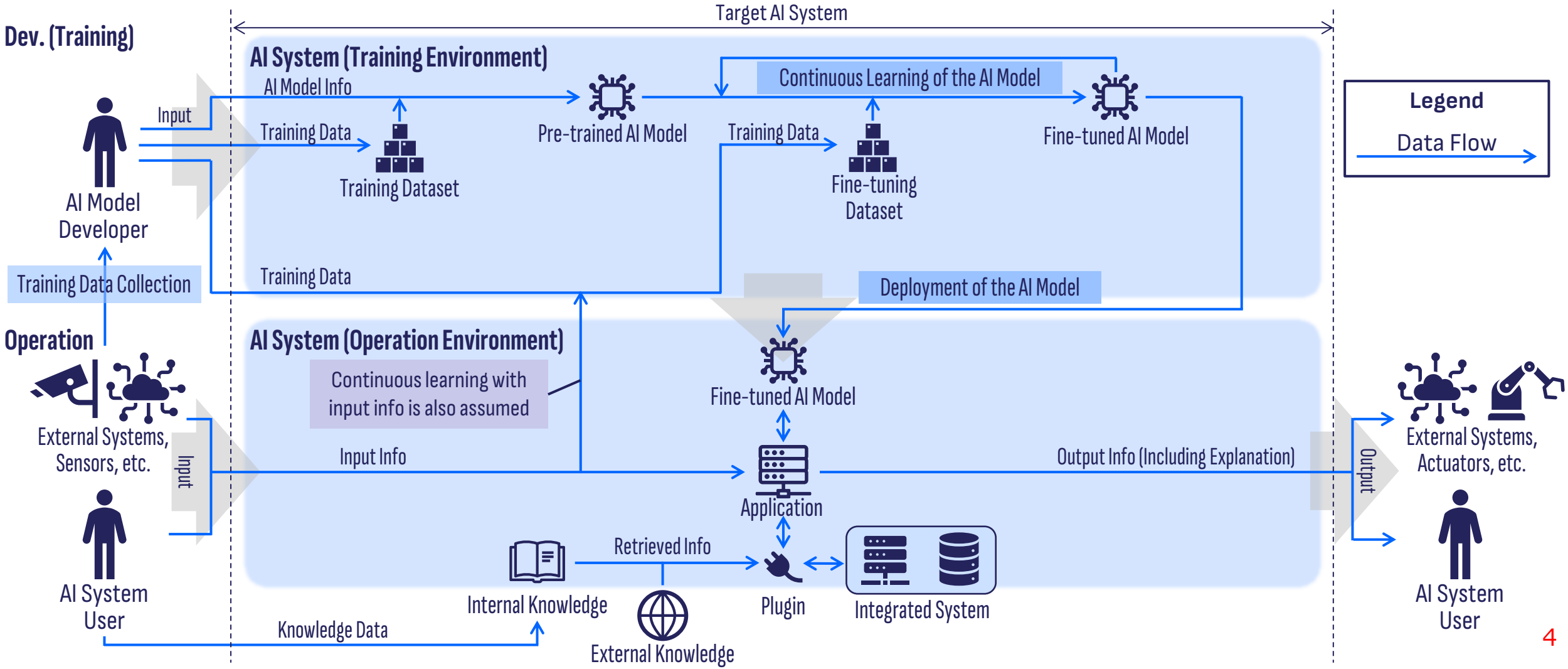
- Measures against conventional cybersecurity attacks are assumed to be necessary.
- This document focuses on known attacks that require special attention when a system incorporates AI, specifically, those involve **intervening in an AI model's training or inference processes**, as well as **exploiting or tampering with its inputs and outputs**.
  - Examples **Within** the Scope:
    - *During Training:* Data Poisoning Attacks (Intervention in the Training Process)
    - *During Inference:* Model Extraction Attacks (Exploit the AI model's output)
  - Examples **Outside** the Scope:
    - Theft of AI Models Through Unauthorized Access by Exploiting Network Device Vulnerabilities.

      Note: Although this attack targets AI models, it does not involve interfering with the training or inference process of the AI models in the target system, nor does it manipulate its input or output.

AI Systems Assumed in This Document

AISI Japan AI Safety Institute

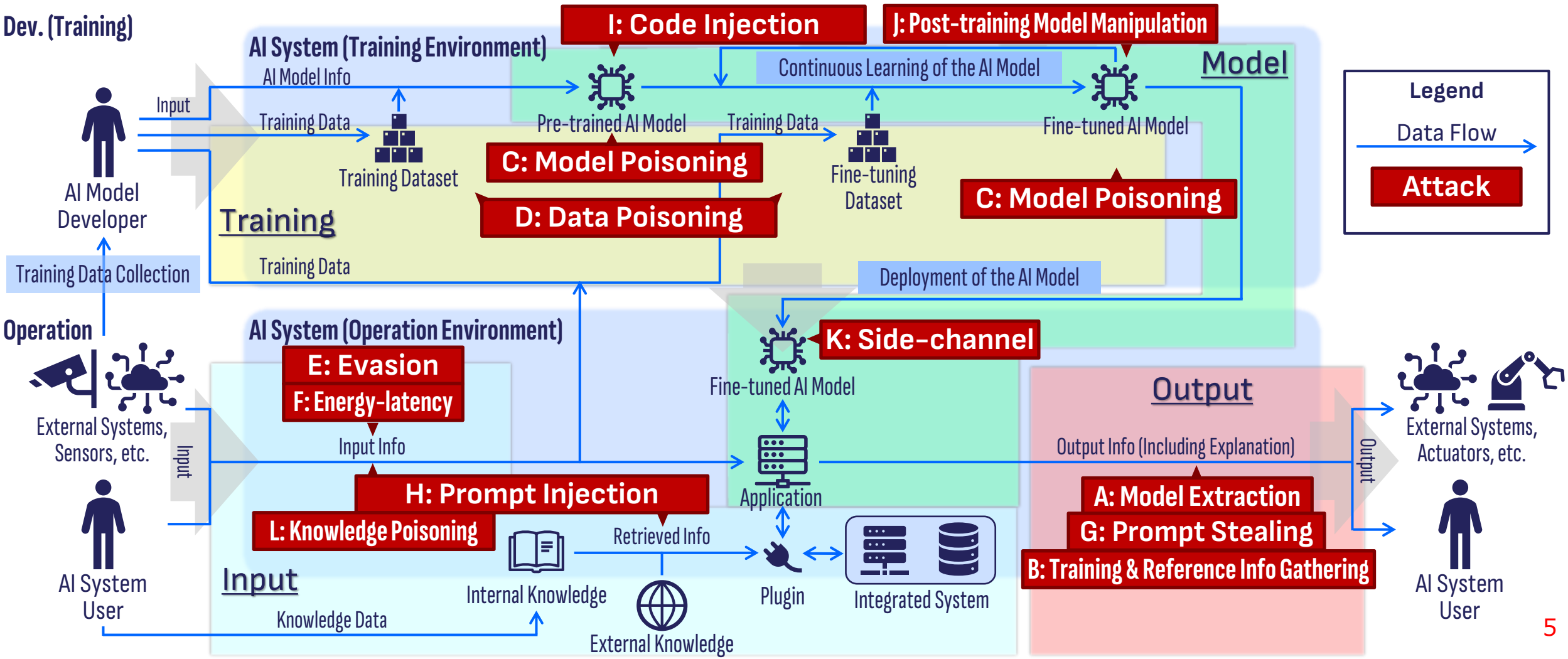
- The assumed AI system consists of two environments: development (training) and operation.
- The AI model is not limited to LLMs, while the figure below includes components often used with LLMs.



Attacks on AI Systems: Overview

AISI Japan AI Safety Institute

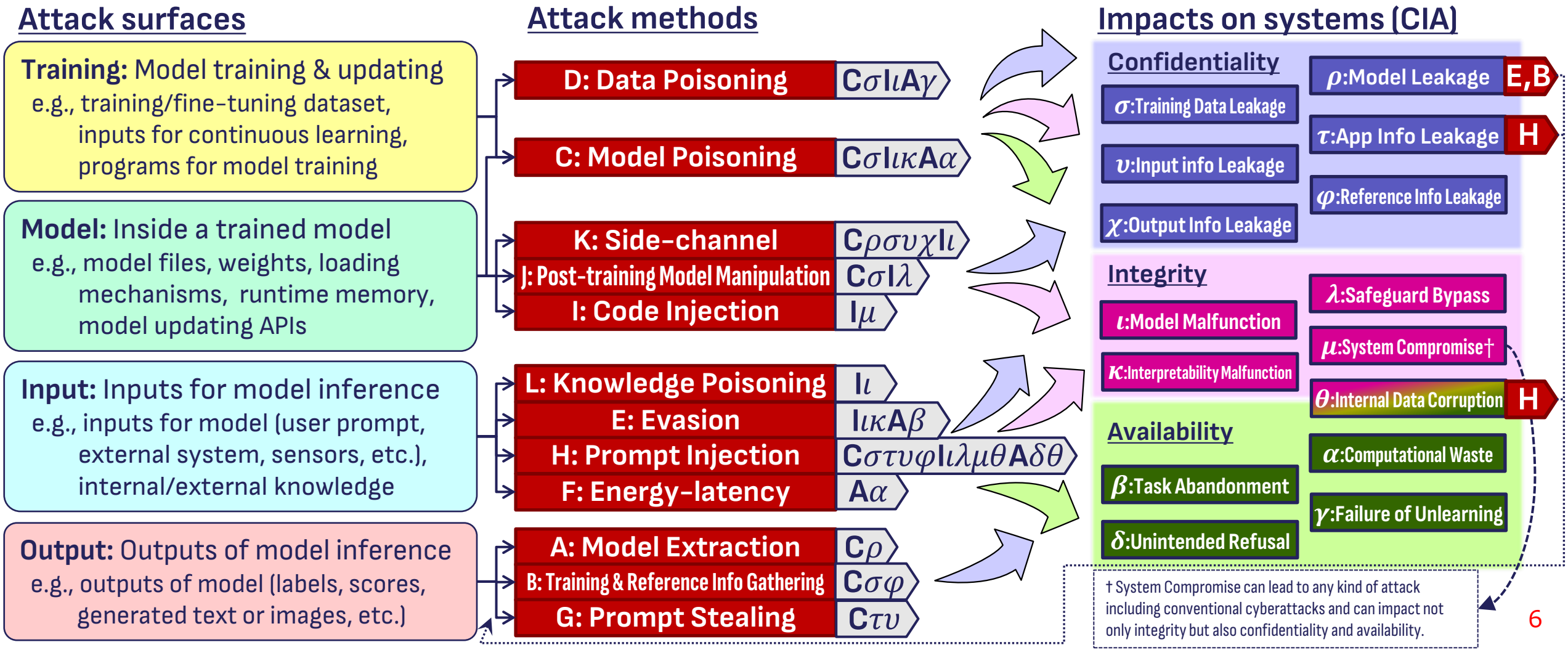
- Attacks are classified into 12 types (A-L) across 4 attack surfaces: Training, Model, Input, and Output.
- Some of these types can be further divided into more specific subtypes.



Attacks and Their Impacts on AI Systems: Overview

AISI Japan AI Safety Institute

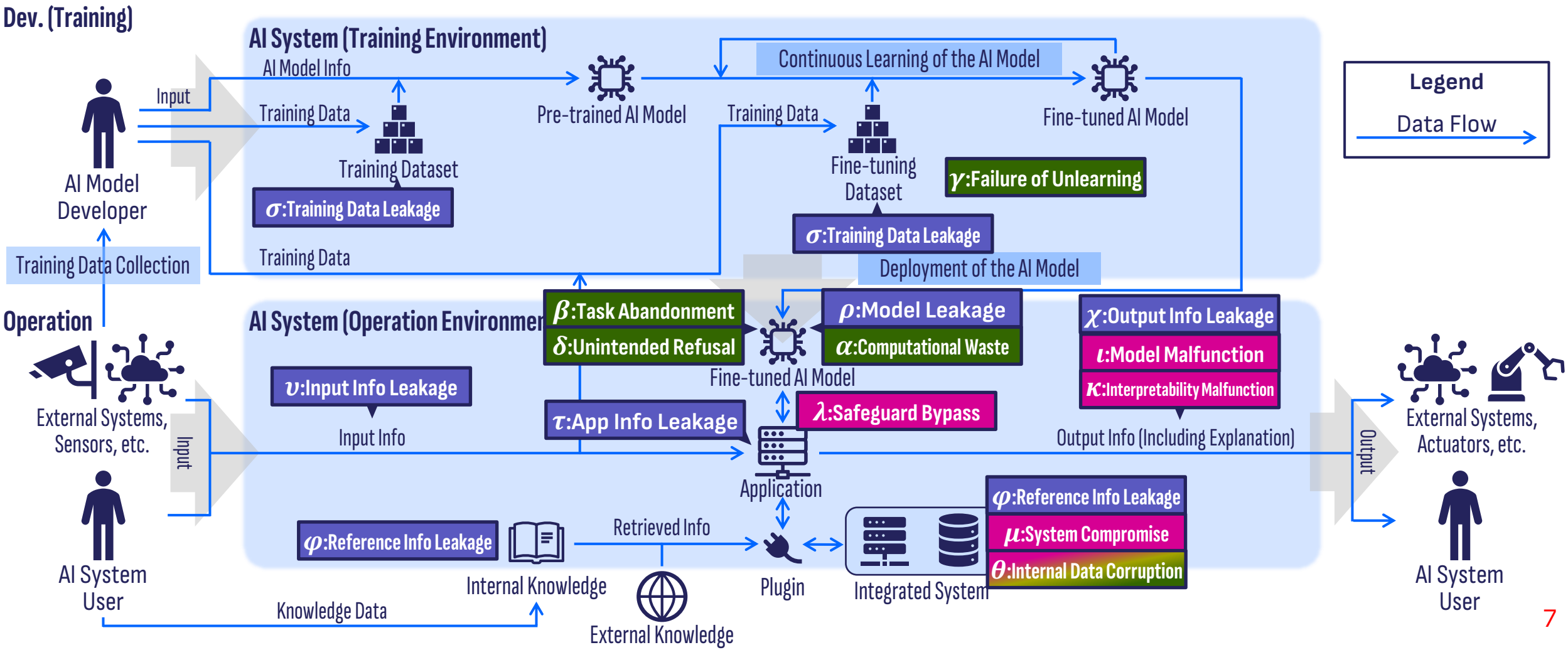
- The figure below illustrates the relationship between attack surfaces, attack methods, and their impacts on systems from the perspective of the security CIA triad.



Impacts of Attacks on AI Systems: Overview

AISI Japan AI Safety Institute

- The figure below illustrates the impacts of attacks on the components of the assumed AI system.
- In some cases, the same type of impact (e.g., Model Leakage) may occur at multiple points.



Attacks and Their Impacts on AI Systems: Detailed Lists 1

AISI Japan AI Safety Institute

Table 1: Attacks Targeting Training as the Primary Attack Surface and Their Impacts

Attack	Impact		Attacker's Capability (Basic Assumptions)	Modality : Target AI[Literature] (This is summary from literature and does not limit modality or target)
C Model Poisoning	C	σ:Training Data Leakage	Training program tampering	Image: NN [SRS17] Text: SVM, LR [SRS17]
	I	ι:Model Malfunction	Malicious adapter distribution	Text: LM [DXC+25]
			Malicious model distribution	Any: NN [LSZ+25]
		κ:Interpretability Malfunction	Model tampering	Tabular: NN, EM [SHJ+20]
	A	α:Computational Waste	Model tampering	Image: NN [CDB+23]
D Data Poisoning	C	σ:Training Data Leakage	Training data injection & Massive queries to the model	Tabular: LR, NN [MGC22] Text: LR [MGC22]
	I	ι:Model Malfunction	Training data injection	Image: NN [Car21], SuperNet [OVA+25] Text: LM [Sch19]
				Tabular: Linear Regression [JOB+18], Text to Image: Diffusion [NRB+25; LPC+25]
			Training data injection & Malicious model distribution for model merge	Image: ViT [WWC+25], Text: LM [WWC+25]
	A	γ:Failure of Unlearning	Training data injection	Image: NN [YZL+25]

Table 2: Attacks Targeting Model as the Primary Attack Surface and Their Impacts

Attack	Impact		Attacker's Capability (Basic Assumptions)	Modality : Target AI[Literature] (This is summary from literature and does not limit modality or target)
I Code Injection	I	μ:System Compromise	Malicious model distribution	Any: Any [ZWZ+24; ZCS+25]
J Post-training Model Manipulation	C	σ:Training Data Leakage	Access to fine-tuning functionality	Text: LM [CTZ+24]
	I	λ:Safeguard Bypass	Access to the model's internal info	Text: LM [YYC+24; KDD25]
			Access to unlearning functionality	Text: LM [SKK+25]
			Access to the fine-tuning functionality	Text to Image: Diffusion [WYB+25]
K Side-channel	C	ρ:Model Leakage	Access to physical memory	Image: NN [RCY+22]
			Physical access to the device	Image: NN on GPU [HCW+25]
		σ:Training Data Leakage	Access to physical memory & Massive queries to the model	Image, Video: NN in TEE [YLD+25]
		υ:Input Info Leakage	Access to physical memory & Massive queries to the model	Image, Video: NN in TEE [YLD+25]
			Share the common MoE model library	Text, Image: MoE [DXS+25]
			Access to the CPU cache	Text: LM [GHG+25]
			Access to hypervisor control	Text: LM [YHG+25]
		χ:Output Info Leakage	Share the common MoE model library	Text, Image: MoE [DXS+25]
			Access to the CPU cache	Text: LM [GHG+25]
	I	ι:Model Malfunction	Access to physical memory	Image: NN [LWX+24; CLY+25]
			Massive queries to the model & Access to physical memory	Image, Text: LM, Transformer [NMF+24]
			Access to GPU memory	Image: NN on GPU [LQS25]

ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

Attacks and Their Impacts on AI Systems: Detailed Lists 2

AISI Japan AI Safety Institute

Table 3: Attacks Targeting Input as the Primary Attack Surface and Their Impacts

Attack		Impact		Attacker's Capability (Basic Assumptions)	Modality : Target AI[Literature] (This is summary from literature and does not limit modality or target)
E	Evasion	I	ι:Model Malfunction	Massive queries to the model	Image: NN, LR, SVM, DT, kNN [PMG+17], Transformer [LHS+25], Text: LM[BSA+22], Image to Text: VLM [WBH+25]
				Access to the model's internal info	Image: SVN [MMR+25], NN [SZS+14;GSS15;MMR+25], Binary file: NN [NFI+25], Text: LSTM [ERL+18], Transformer [MMR+25], A specific text similarity matching algorithm [HRS25], Weather data: Diffusion [IER25]
				Input to the model	Text: Transformer [GSJ+21], LSTM, DA [WFK+19], LiDAR Point Cloud: NN [KNT+25], Image: NN [LLM+25;XDT+25], CV [XC25], Multi-modal (Image & Text) : VLM [WZS+25], Image to Text: VLM [XDT+25], Audio to Text: ASR [YZG+25]
				Malicious input distribution	Text: LM [LLW+25b]
				Transmit adversarial signal	Radio Signals: JSCC [CSH+25]
			κ:Interpretability Malfunction	Massive queries to the model	Image: NN[DAA+19]
		A	β:Task Abandonment	Access to the model's internal info	Audio: Transformer [WLC+25]
F	Energy-latency	A	α:Computational Waste	Input to the model	Image・Text: NN [SZB+21]
				Massive queries to the model	Text: LM [BSA+22]
H	Prompt Injection	C	σ:Training Data Leakage	Input to the model	Text: LM [KKC25]
			τ:App Info Leakage	Input to the model	Text: LM[ZCI24]
			υ:Input Info Leakage	User-referenced info poisoning	Text: A specific chat service [Sam23]
			φ:Reference Info Leakage	Input to the model	Text: LM [CBN25], Text to Tabular: LM [PEC+25]
		I	ι:Model Malfunction	Malicious image distribution	Multi-modal (Image & Text): VLM [ZZM+25]
			λ:Safeguard Bypass	Input to the model	Text: LM [LDL+24;SCB+24;ZWC+23;WHS23;CRD+24;MZK+24;PHS+22;WFK+19;RSE25], Text to Image: Diffusion [LMX+25;DMY+25], VLM [RSE25], A specific image generation system [VMP25]
				Massive queries to the model	Text: LM [ZGY+25]
				Access to the model's internal info	Text・Image: NN [CNC+23]
				Access to fine-tuning functionality	Text: LM [LPH+25]
			μ:System Compromise	Reference data injection	Text, Image to Text: LM [GAM+23]
		I/A	θ:Internal Data Corruption	Input to the model	Text to Tabular: LM [PEC+25]
		A	δ:Unintended Refusal	Reference data injection	Text: LM [SSS25]
L	Knowledge Poisoning	I	ι:Model Malfunction	Reference data injection	Text: LM[YSQ25;BS25;GCL+25;ZGW+25]
				Reference data injection & Massive queries to the model	Text: LM[CGL+25]

ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

Attacks and Their Impacts on AI Systems: Detailed Lists 3

AISI Japan AI Safety Institute

Table 4: Attacks Targeting Output as the Primary Attack Surface and Their Impacts

Attack			Impact		Attacker's Capability (Basic Assumptions)	Modality : Target AI[Literature] (This is summary from literature and does not limit modality or target)
A	Model Extraction	C	ρ:Model Leakage		Massive queries to the model	Image: NN [OSF19a;OSF19b;JSM+19;CBB+18;YZW+25], LR[TZJ+16] Tabular: RR[WG18], DT[TZJ+16], LR, NN, SVM[TZJ+16,WG18], Text: LSTM [YZW+25]
B	Training & Reference Info Gathering: Membership Inference	C	σ:Training Data Leakage		Input to the model	Image: RR, DT [YGF+18], NN [AYM+19;YGF+18;LLW+25a], Tabular: RR, DT [YGF+18], NN [AYM+19;YGF+18], MLP [LLW+25a], Text: NN [AYM+19], LM [HLL+25a], MLP [LLW+25a], Image: Diffusion [PW25]
					Massive queries to the model	Image: NN [SSS+17;LBW+18;YZW+25;NYW+25], ViT [NYW+25] Tabular: NN [SSS+17;LBW+18], Text: LM [LSS+23;NYW+25], LSTM [YZW+25], Multi-modal (Image & Text): VLM [HLL+25b]
					Access to the model's internal info	Image, Tabular: NN [NSH19]
					Input to the model & Malicious model distribution	Image: NN [WAA+25]
			φ:Reference Info Leakage		Input to the model	Text: LM [NPS+25;GMD+25]
	Attribute Inference	C	σ:Training Data Leakage		Input to the model	Image, Tabular: LR, DT [YGF+18]
					Access to the model's internal info	Image, Text: NN [SS20]
					Massive queries to the model	Tabular: NN [KCM25]
	Property Inference	C	σ:Training Data Leakage		Access to the model's internal info	Image, Tabular: NN [GWY+18] Audio, Network traffic: NN, SVM, HMM, DT[AMS+15]
	Model Inversion	C	σ:Training Data Leakage		Massive queries to the model	Image: DT [FJR15], NN[FJR15;YZW+25], Tabular: DT, NN [FJR15], Text: LSTM [YZW+25]
					Access to the model's internal info	Text: Transformer[ZHK22]
	Data Reconstruction	C	σ:Training Data Leakage		Access to the model's internal info	Image: NN [BCH22;HVY+22;BHY+23]
					Massive queries to the model	Text: LM [LSS+23]
	Data Extraction	C	σ:Training Data Leakage		Massive queries to the model	Text: LSTM, RNN [CLE+19], LM [CTW+21;LSS+23;LWW+25], Text to Image: Diffusion [CHN+23]
					Input to the model	Text: LM [CMX+25;KSM+25]
G	Prompt Stealing	C	τ:App Info Leakage		Input to the model	Text: LM [YLL+25]
			υ:Input Info Leakage		Access to the model's output	Text to Image: Diffusion [SQB+24]
					Input to the model	Text: LM [TSS+25]
					Massive queries to the model	Text: LM [YZH+25], Text to Image: Diffusion [YZH+25]

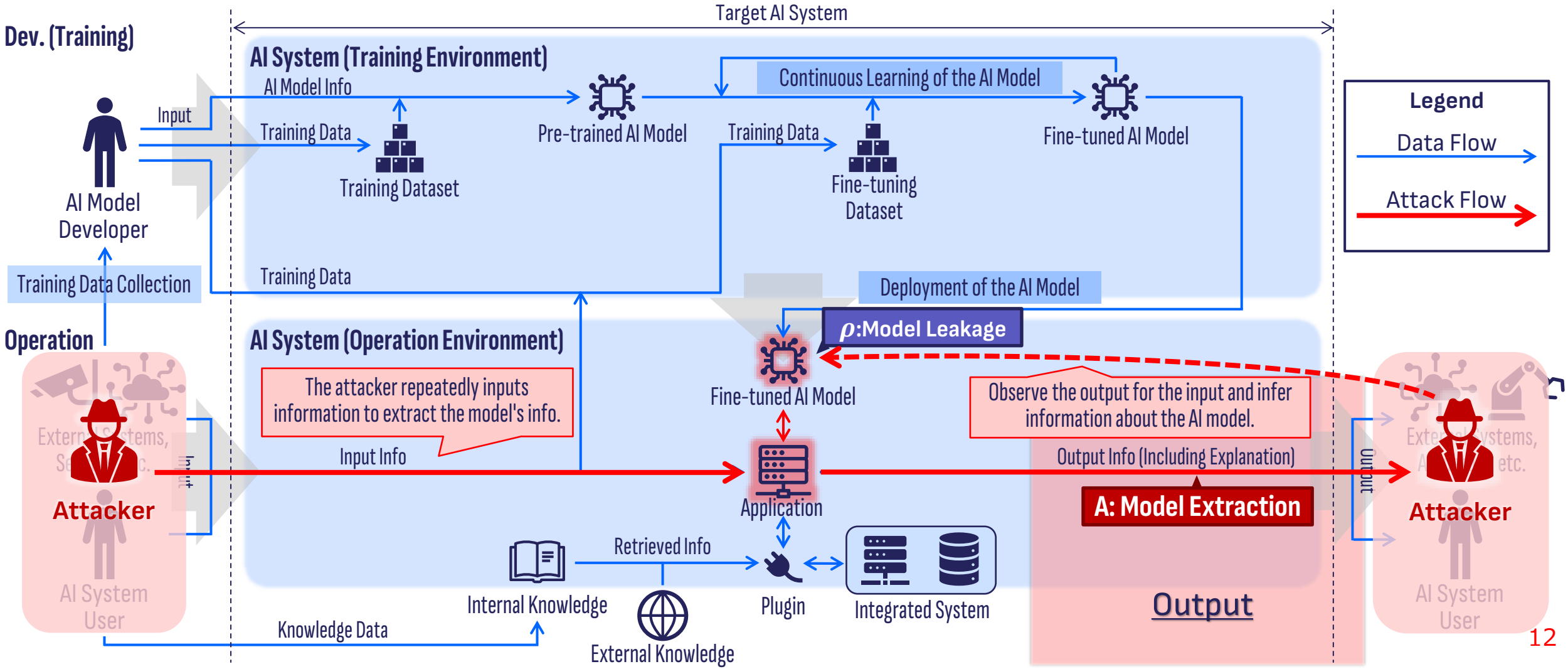
ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

General Descriptions of Each Attack

Attack A: Model Extraction

AISI Japan AI Safety Institute

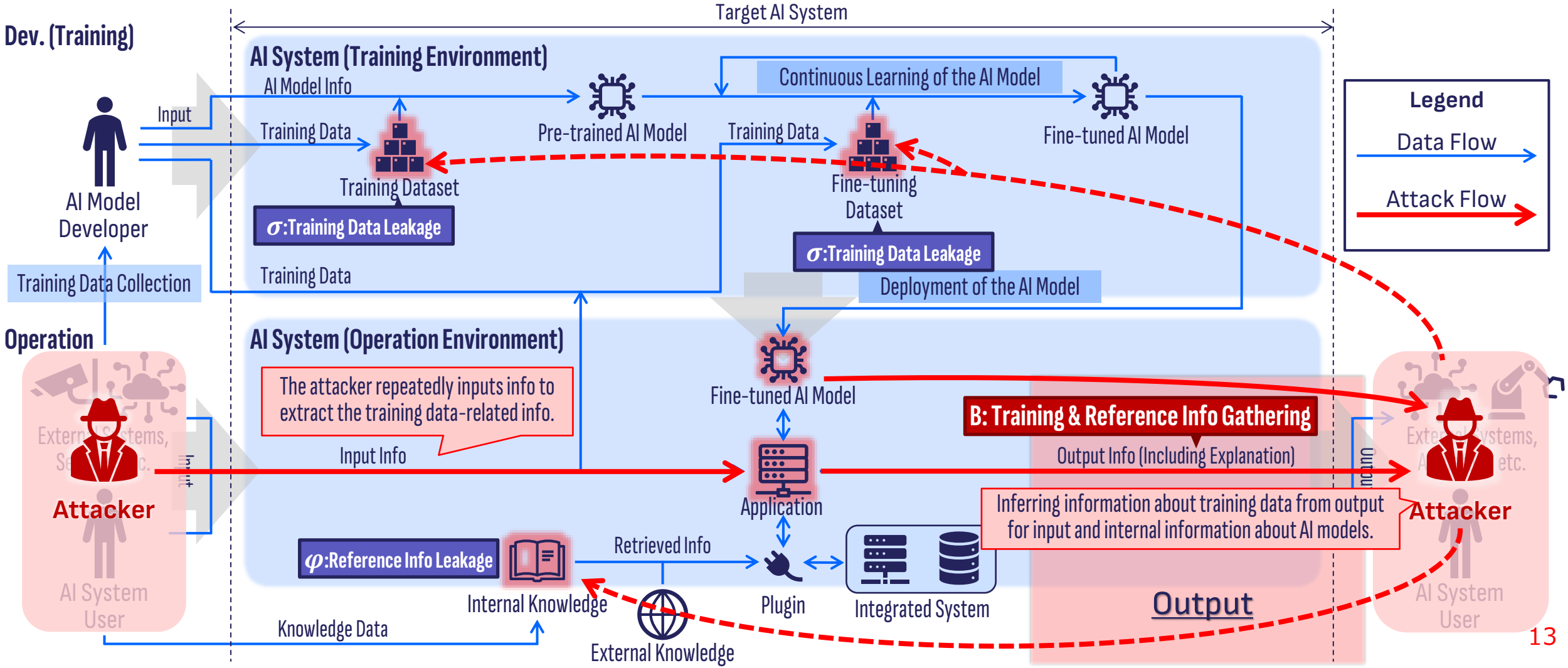
- By observing the output produced in response to given input data, this attack can cause the leakage of information about the AI model without direct access to the model itself.



Attack B: Training & Reference Info Gathering

AISI Japan AI Safety Institute

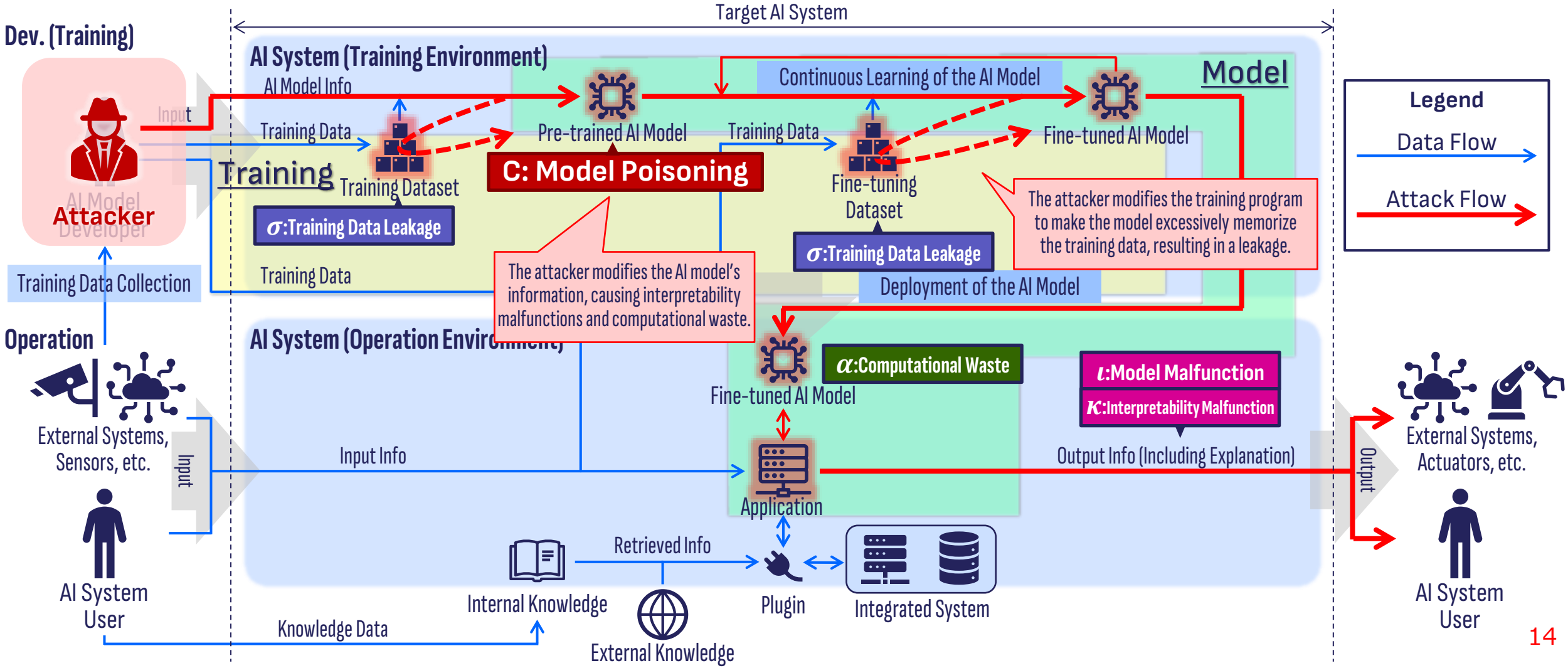
- Information about the training dataset and internal knowledge is primarily obtained through the relationship between its inputs and outputs. There are various attack subtypes targeting different types of information.



Attack C: Model Poisoning

AISI Japan AI Safety Institute

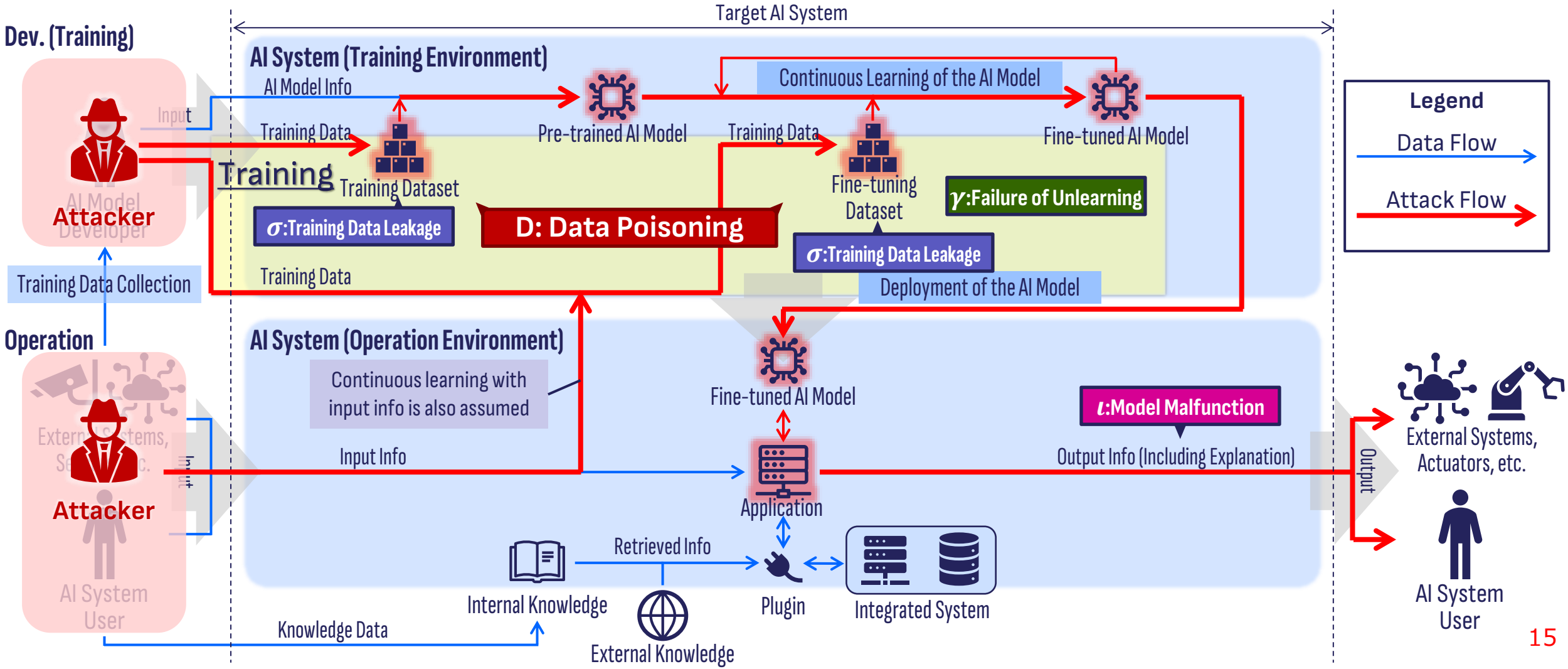
- By modifying the information of the AI model or training programs, this attack can cause interpretability or model malfunctions, computational waste, and training data leakage when the model is in operation.



Attack D: Data Poisoning

AISI Japan AI Safety Institute

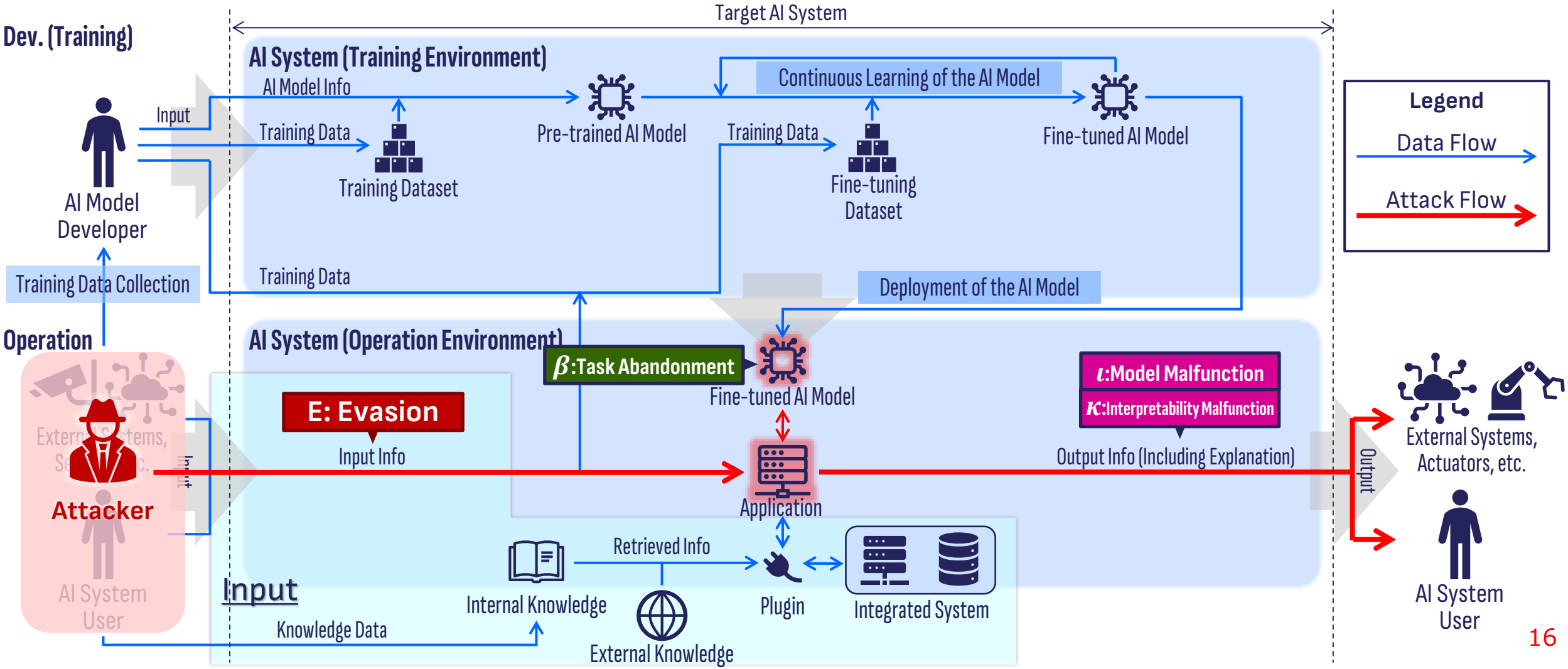
- Adding special data to the training dataset can cause information leakage about the remaining training data, the model malfunction, and the failure of unlearning (the deletion of specific training samples).



Attack E: Evasion

AISI Japan AI Safety Institute

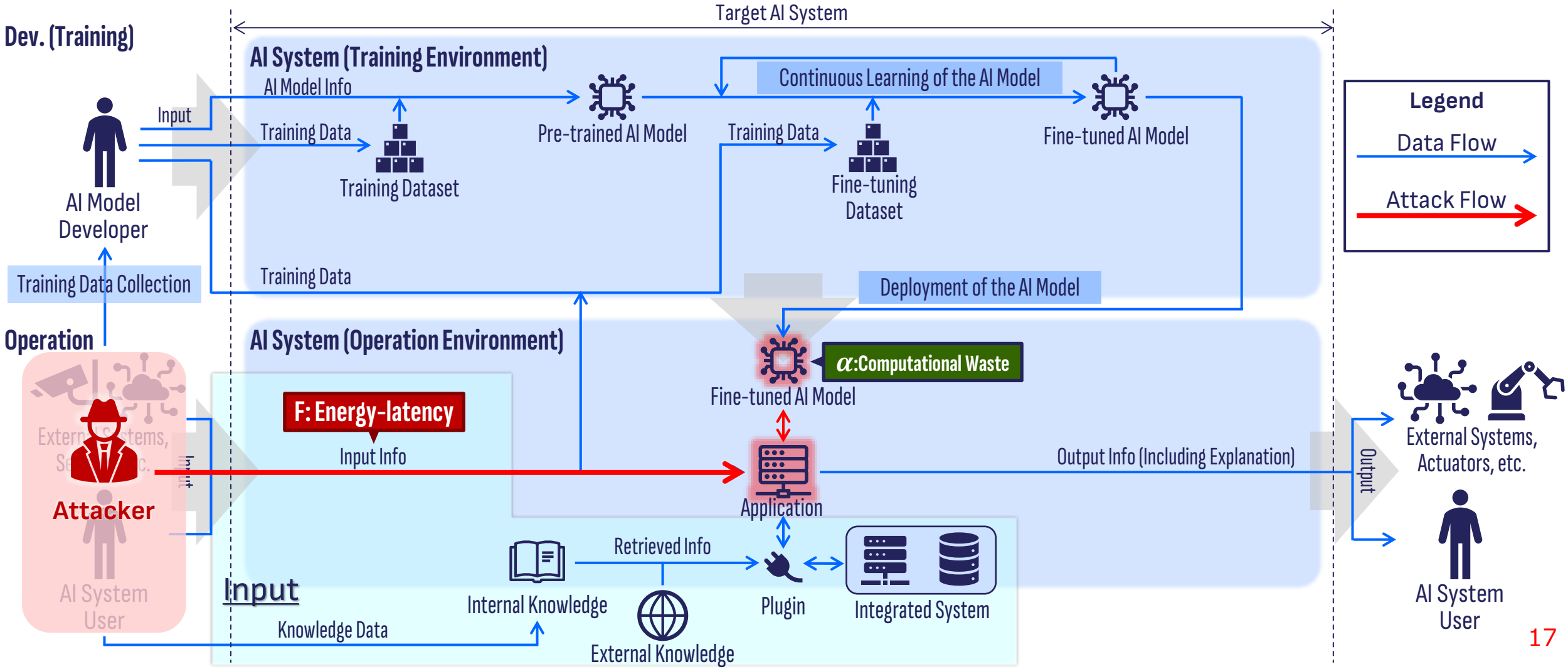
- Adversarial examples (specially crafted inputs for AI systems) can cause task abandonment and malfunctions in both output and interpretability when provided during operation.



Attack F: Energy-latency

AISI Japan AI Safety Institute

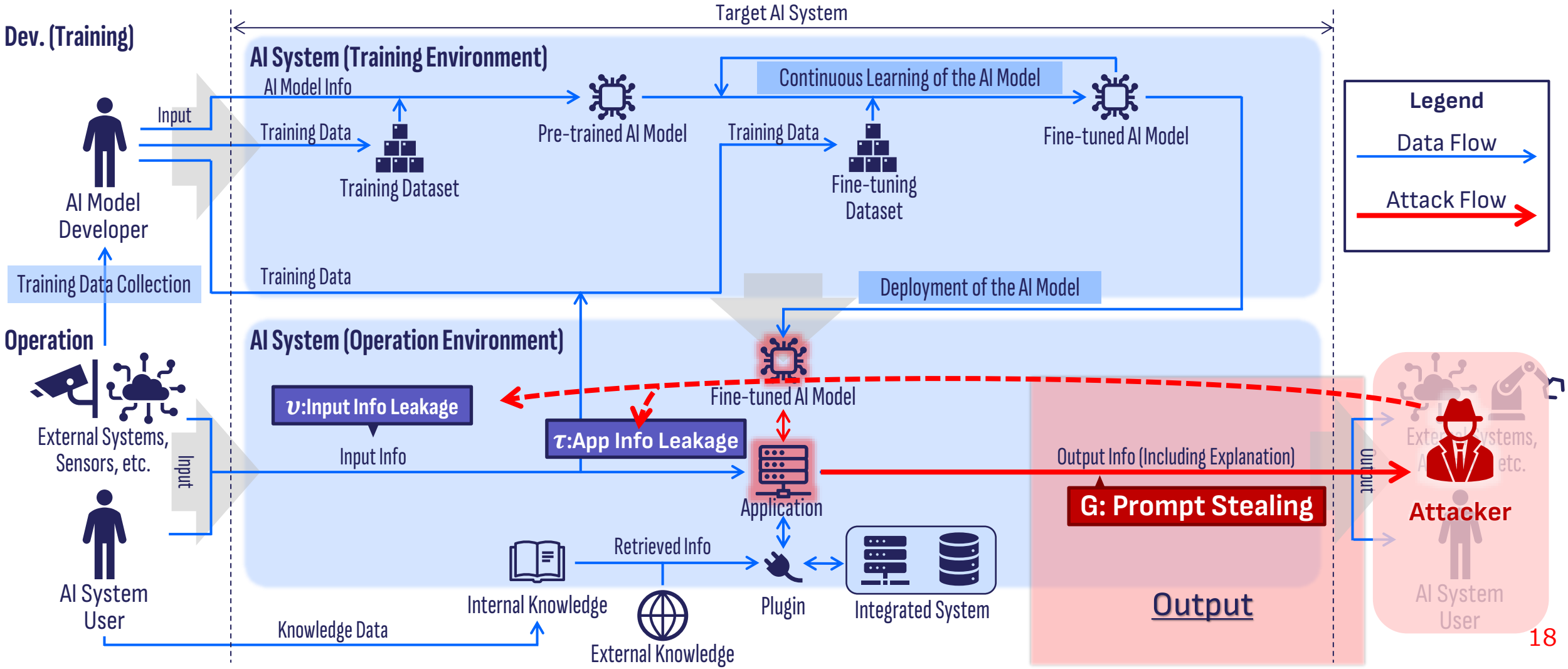
- By providing the AI system with specially crafted inputs called sponge examples during operation, this attack can waste computing resources, such as by increasing response time and energy consumption.



Attack G: Prompt Stealing

AISI Japan AI Safety Institute

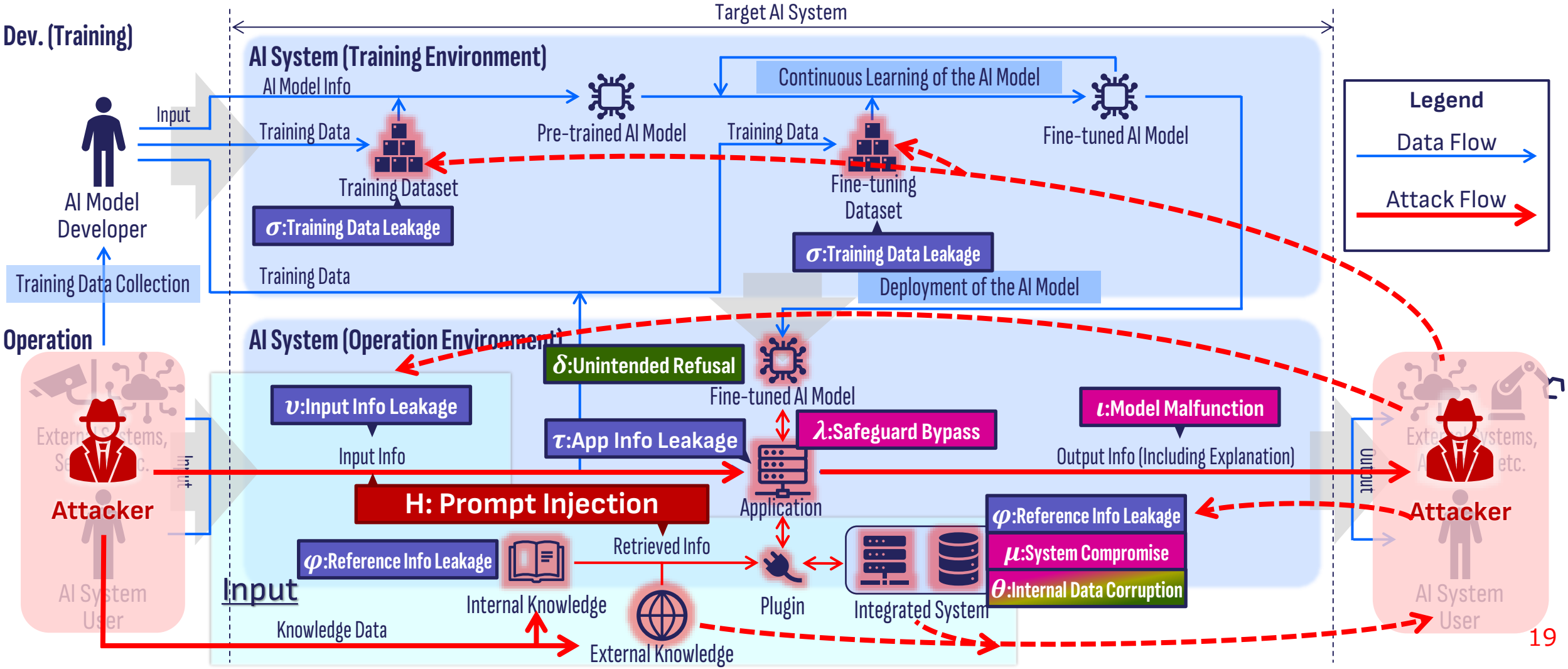
- By inferring the original prompt from outputs such as images generated by the AI model, this attack can leak information about input and application that constitutes prompt engineering know-how.



Attack H: Prompt Injection

AISI Japan AI Safety Institute

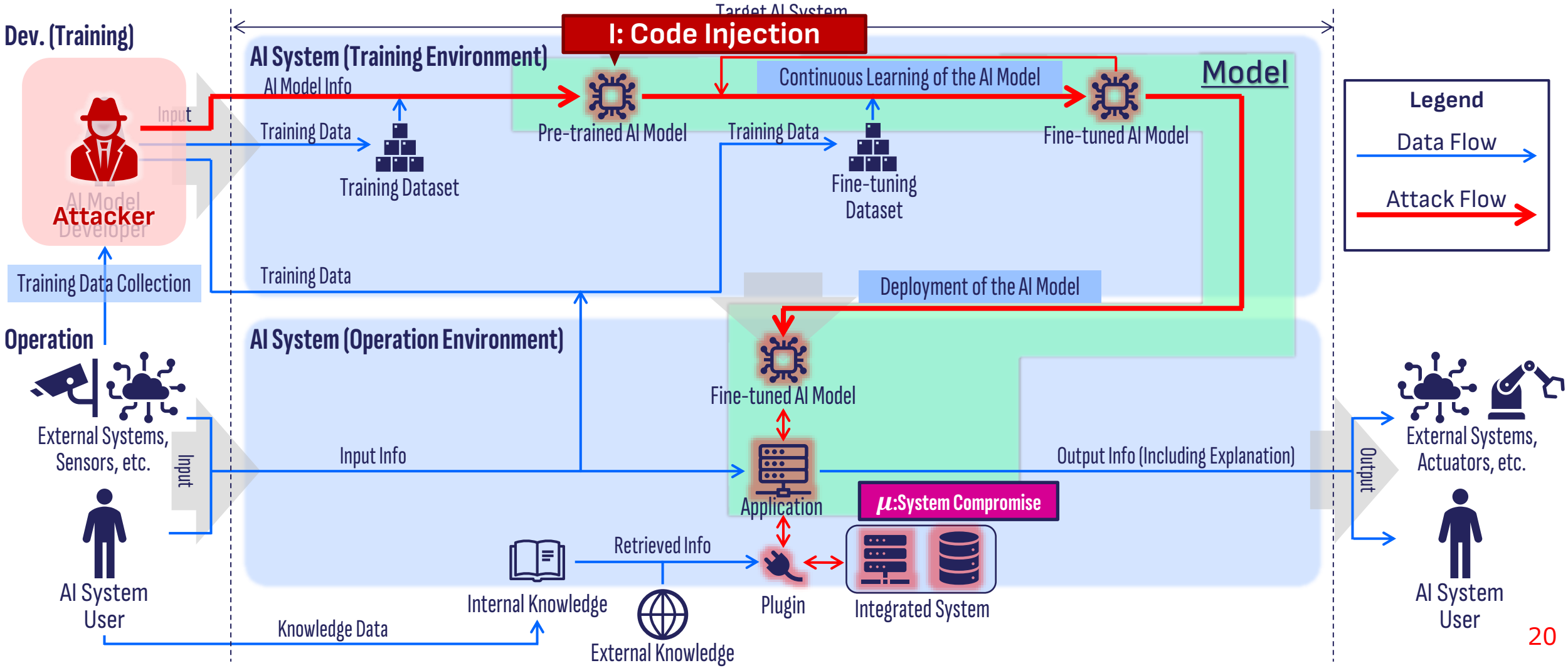
- By directly inputting adversarial instructions into the model or indirectly injecting them through its information sources, this attack can cause various forms of information leakage, safeguard bypass, and system compromise.



Attack I: Code Injection

AISI Japan AI Safety Institute

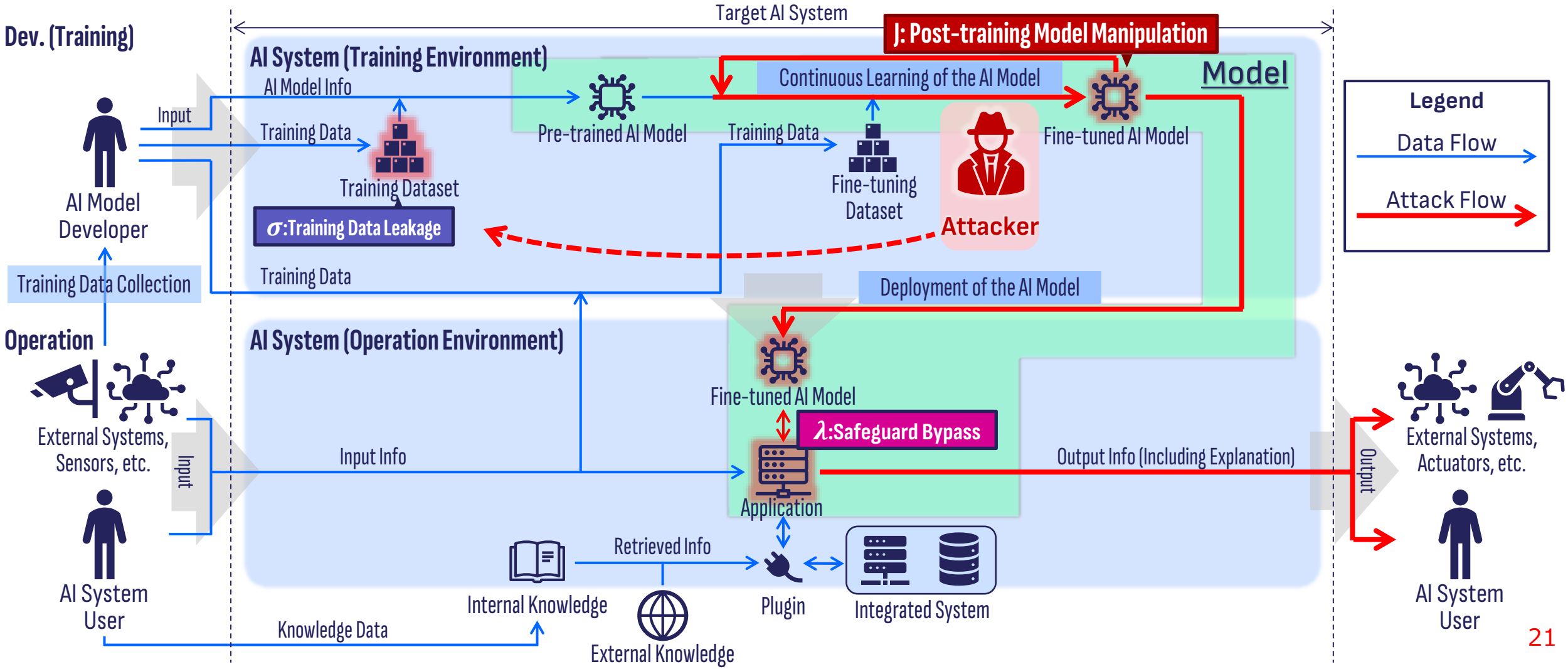
- Embedding executable code within the AI model that is then executed when the model is invoked can compromise the system.



Attack J: Post-training Model Manipulation

AISI Japan AI Safety Institute

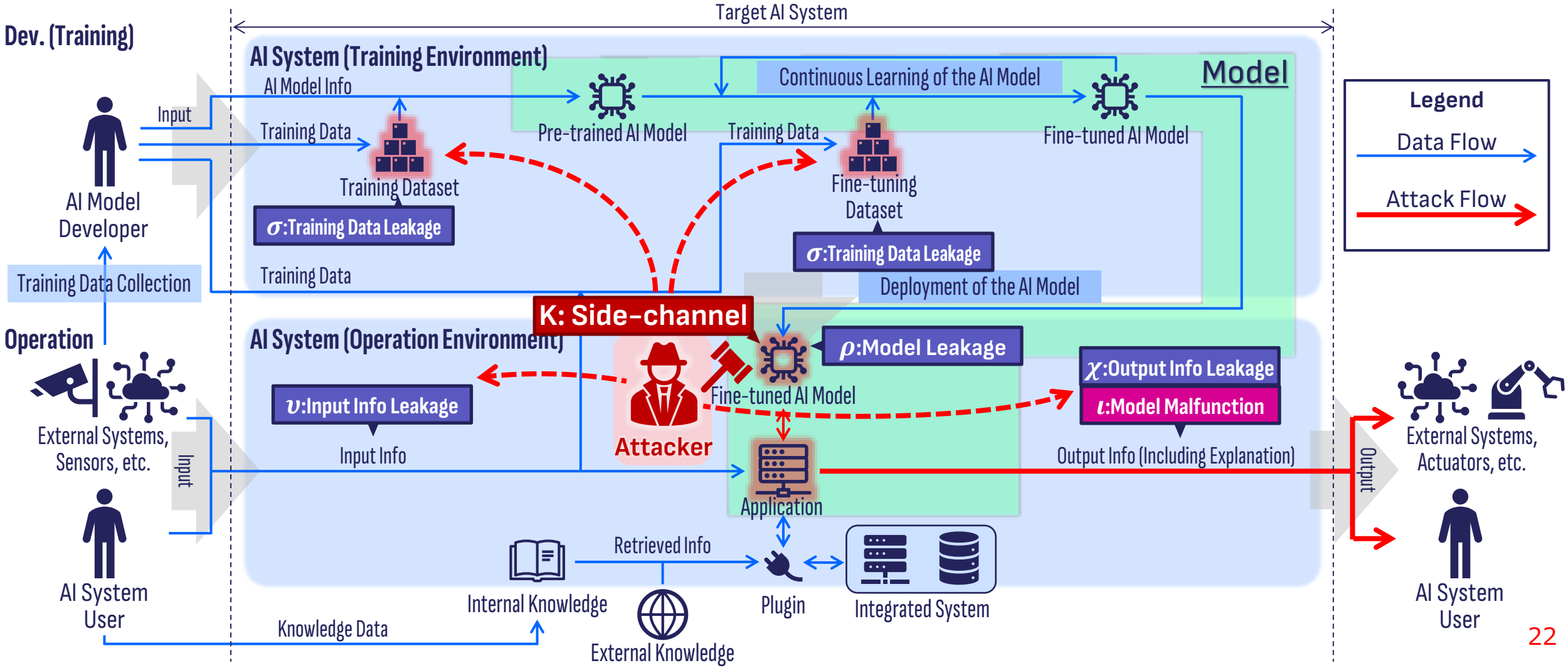
- By performing harmful fine-tuning, unlearning, and pruning, the target pre-trained AI model is manipulated to bypass safeguards and leak its pre-trained data.



Attack K: Side-channel

AISI Japan AI Safety Institute

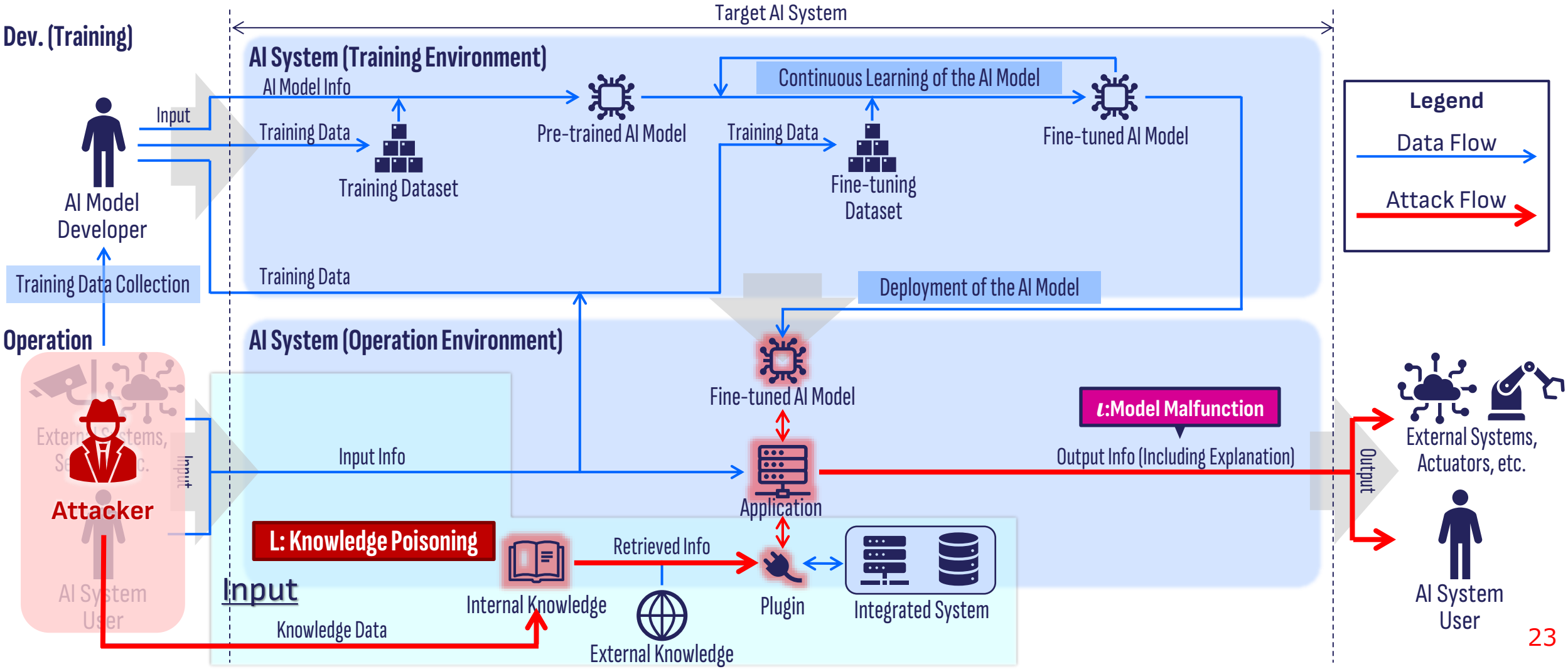
- Attackers who have access to operational environments, such as cloud-based systems and edge devices, can cause information leaks and model malfunctions by conducting physical surveillance or interference.



Attack L: Knowledge Poisoning

AISI Japan AI Safety Institute

- By injecting specific data into internal knowledge, such as RAG, from an external source, this attack can cause the AI model to malfunction.



Update History

Summary of Updates for the 2nd Edition (Apr. 2026)

AISI Japan AI Safety Institute

- Since the 1st edition was published (March 2025), we have reviewed papers accepted at major international conferences* held in 2025 to incorporate the latest attacks into this document and have included 64 key papers on attacks targeting AI systems.
  *Target conferences: USENIX Security, ACM CCS, IEEE S&P, NDSS
- Add and Rename Attacks / Organize on the Attack Surfaces (See page 26 for details)
  - Added 2 new attack types and renamed 3 existing ones based on the survey results.
  - Introduced a classification based on attack surfaces to organize attacks.
- Add and Rename Impacts of Attacks (See page 27 for details)
  - Added 4 new impacts and renamed 1 existing one based on the survey results.
  - Color-coded visualization of impacts by CIA triad.

Add and Rename Attacks / Organize on the Attack Surfaces

AISI Japan AI Safety Institute

1st Edition (Mar. 2025)

- A: Model Extraction
- B: Training Data-related Info Gathering
  - Membership Inference
  - Attribute Inference
  - Property Inference
  - Model Inversion
  - Data Extraction
- C: Model Poisoning
- D: Data Poisoning
- E: Evasion
- F: Energy-latency
- G: Prompt Stealing
- H: Prompt Injection
- I: Code Injection
- J: Adversarial Fine-tuning
- K: Rowhammer

2nd Edition (Apr. 2026)

- A: Model Extraction
- B: Training & Reference Info Gathering — Rename
  - Membership Inference
  - Attribute Inference
  - Property Inference
  - Model Inversion
  - Data Reconstruction — Add
  - Data Extraction
- C: Model Poisoning
- D: Data Poisoning
- E: Evasion
- F: Energy-latency
- G: Prompt Stealing
- H: Prompt Injection
- I: Code Injection
- J: Post-training Model Manipulation — Rename
- K: Side-channel — Rename
- L: Knowledge Poisoning — Add

Attack surfaces

Training: Model training & updating
e.g., training/fine-tuning dataset, inputs for continuous learning, programs for model training

Model: Inside a trained model
e.g., model files, weights, loading mechanisms, runtime memory, model updating APIs

Input: Inputs for model inference
e.g., inputs for model (user prompt, external system, sensors, etc.), internal/external knowledge

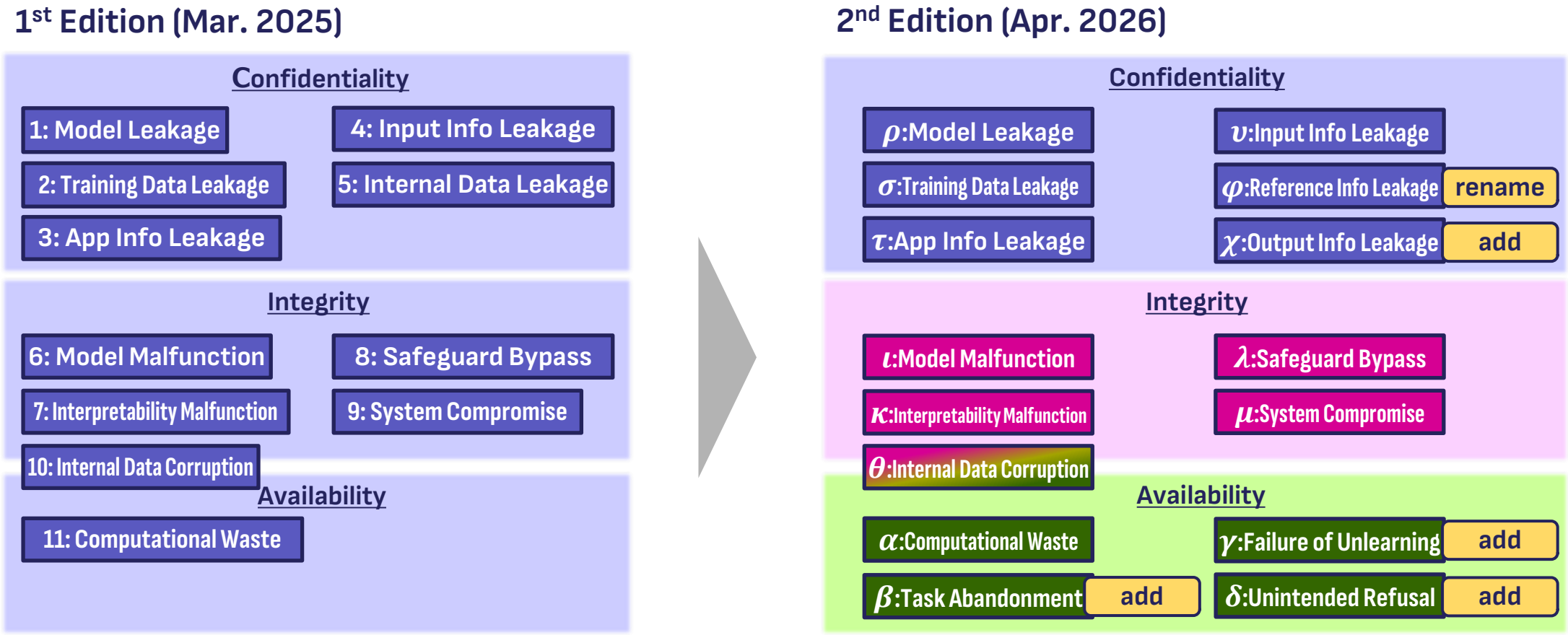
Output: Outputs of model inference
e.g., outputs of model (labels, scores, generated text or images, etc.)

The arrow above shows the main relationship.

Add and Rename Impacts of Attacks

AISI Japan AI Safety Institute

- We added new types of impacts and renamed some of them. We also changed the labels for each impact from numbers to Greek letters and color-coded them by CIA triad for better visualization.



References (Part 1 of 10)

AISI Japan AI Safety Institute

[AMS+15] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, et al. “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers”. In: Int. J. Secur. Netw. 10.3 (Sept. 2015), pp. 137—150. doi: 10.1504/IJSN.2015.071829.

[AYM+19] Salem Ahmed, Zhang Yang, Humbert Mathias, Berrang Pascal, Fritz Mario, et al. “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models”. In: Proceedings 2019 Network and Distributed System Security Symposium (2019). doi: 10.14722/ndss.2019.23119.

[BHY+23] Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, and Michal Irani. Reconstructing Training Data from Multiclass Neural Networks. 2023.

[BS25] Matan Ben-Tov and Mahmood Sharif. “GASLITEing the Retrieval: Exploring Vulnerabilities in Dense Embedding-based Search”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4364—4378. doi: 10.1145/3719027.3765095.

[BSA+22] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. “Bad Characters: Imperceptible NLP Attacks”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1987—2004. doi: 10.1109/SP46214.2022.9833641.

[Car21] Nicholas Carlini. “Poisoning the Unlabeled Dataset of Semi-Supervised Learning”. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 1577—1592. url: https: //www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

[CBB+18] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. “Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data”. In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018, pp. 1—8. doi: 10.1109/IJCNN.2018.8489592.

[CBN25] Stav Cohen, Ron Bitton, and Ben Nassi. “Here Comes the AI Worm: Preventing the Propagation of Adversarial Self-Replicating Prompts Within GenAI Ecosystems”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3975—3989. doi: 10.1145/3719027.3765196.

[CDB+23] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-Latency Attacks via Sponge Poisoning. 2023. arXiv: 2203.08147 [cs.CR].

[CGL+25] Zhuo Chen, Yuyang Gong, Jiawei Liu, Miaokun Chen, Haotan Liu, et al. “FlippedRAG: Black-Box Opinion Manipulation Adversarial Attacks to Retrieval-Augmented Generation Models”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4109—4123. doi: 10.1145/3719027.3765023.

[CHN+23] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, et al. “Extracting Training Data from Diffusion Models”. In: 32nd USENIX Security Symposium (USENIX Security 23). Anaheim, CA: USENIX Association, Aug. 2023, pp. 5253—5270. url: https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks”. In: 28th USENIX Security Symposium (USENIX Security 19). Santa Clara, CA: USENIX Association, Aug. 2019, pp. 267—284. url: https: //www.usenix.org/conference/usenixsecurity19/presentation/carlini.

[CLY+25] Yanzuo Chen, Zhibo Liu, Yuanyuan Yuan, Sihang Hu, Tianxiang Li, et al. “Compiled Models, Built-In Exploits: Uncovering Pervasive Bit-Flip Attack Surfaces in DNN Executables”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, California, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/compiled-models-built-inexploits-uncovering-pervasive-bit-flip-attack-surfaces-in-dnn-executables/.

References (Part 2 of 10)

AISI Japan AI Safety Institute

[CMX+25] Shuai Cheng, Shu Meng, Haitao Xu, Haoran Zhang, Shuai Hao, et al. “Effective PII Extraction from LLMs through Augmented Few-Shot Learning”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/cheng-shuai.

[CNC+23] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, et al. “Are aligned neural networks adversarially aligned?” In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 61478—61500. url: https://proceedings.neurips.cc/paper_files/paper/2023/hash/ c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

[CRD+24] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, et al. Jailbreaking Black Box Large Language Models in Twenty Queries. 2024. arXiv: 2310.08419 [cs.LG].

[CSH+25] Jung-Woo Chang, Ke Sun, Nasimeh Heydaribeni, Seira Hidano, Xinyu Zhang, et al. “Magmaw: Modality-Agnostic Adversarial Attacks on Machine Learning-Based Wireless Communication Systems”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/magmaw-modalityagnostic-adversarial-attacks-on-machine-learning-based-wireless-communication-systems/.

[CTW+21] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, et al. “Extracting Training Data from Large Language Models”. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 2633—2650. url: https://www.usenix.org/conference/ usenixsecurity21/presentation/carlini-extracting.

[CTZ+24] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, et al. “The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks”. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS ’24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 1285—1299. doi: 10.1145/3658644.3690325.

[DAA+19] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, et al. “Explanations can be manipulated and geometry is to blame”. In: Advances in Neural Information Processing Systems. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, et al. Vol. 32. Curran Associates, Inc., 2019. url: https://proceedings.neurips.cc/paper/ 2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html.

[DMY+25] Yingkai Dong, Xiangtao Meng, Ning Yu, Zheng Li, and Shanqing Guo. “Fuzz-Testing Meets LLM-Based Agents: An Automated and Efficient Framework for Jailbreaking Text-to-Image Generation Models”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 373—391. doi: 10.1109/SP61157.2025.00119.

[DXC+25] Tian Dong, Minhui Xue, Guoxing Chen, Rayne Holland, Yan Meng, et al. “The Philosopher’s Stone: Trojaning Plugins of Large Language Models”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. Reston, VA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/the-philosophers-stone-trojaning-plugins-of-large-languagemodels/.

[DXS+25] Ruyi Ding, Tianhong Xu, Xinyi Shen, Aidong Adam Ding, and Yunsi Fei. “MoEcho: Exploiting Side-Channel Attacks to Compromise User Privacy in Mixture-of-Experts LLMs”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2159—2173. doi: 10.1145/3719027.3765174.

[ERL+18] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. “HotFlip: White-Box Adversarial Examples for Text Classification”. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia, July 2018, pp. 31—36. doi: 10.18653/v1/P18-2006.

References (Part 3 of 10)

AISI Japan AI Safety Institute

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS ’15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322—1333. doi: 10.1145/2810103.2813677.

[GAM+23] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, et al. “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection”. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. AISec ’23. Copenhagen, Denmark: Association for Computing Machinery, 2023, pp. 79—90. doi: 10.1145/3605764. 3623985.

[GCL+25] Yuyang Gong, Zhuo Chen, Jiawei Liu, Miaokun Chen, Fengchang Yu, et al. “Topic-FlipRAG: Topic-Orientated Adversarial Opinion Manipulation Attacks to Retrieval-Augmented Generation Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/gong-yuyang.

[GHG+25] Zibo Gao, Junjie Hu, Feng Guo, Yixin Zhang, Yinglong Han, et al. “I Know What You Said: Unveiling Hardware Cache Side-Channels in Local Large Language Model Inference”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/gao-zibo.

[GMD+25] Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. “DCMI: A Differential Calibration Membership Inference Attack Against Retrieval-Augmented Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4184—4198. doi: 10.1145/3719027.3765103.

[GSJ+21] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. “Gradient-based Adversarial Attacks against Text Transformers”. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 5747—5757. doi: 10.18653/v1/2021. emnlp-main.464.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2015. arXiv: 1412.6572 [stat.ML].

[GWY+18] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. “Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations”. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. CCS ’18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 619—633. doi: 10.1145/3243734.3243834.

[HCW+25] Peter Horvath, Lukasz Chmielewski, L´eo Weissbart, Lejla Batina, and Yuval Yarom. “BarraCUDA: Edge GPUs do Leak DNN Weights”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/horvath.

[HLL+25a] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, et al. “Towards Label-Only Membership Inference Attack against Pre-trained Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/he-yu.

[HLL+25b] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, et al. “Membership Inference Attacks Against Vision-Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/hu-yuke.

References (Part 4 of 10)

[HRS25] Jhih-Yi (Janet) Hsieh, Aditi Raghunathan, and Nihar B. Shah. “Vulnerability of Text-Matching in ML/AI Conference Reviewer Assignments to Collusions”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/hsieh.

[HVY+22] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. “Reconstructing Training Data from Trained Neural Networks”. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., 2022. url: https://proceedings.neurips.cc/paper_files/paper/2022/file/906927370cbeb537781100623cca6fa6-Paper-Conference.pdf.

[IER25] Erik Imgrund, Thorsten Eisenhofer, and Konrad Rieck. “Adversarial Observations in Weather Forecasting”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3579–3590. doi: 10.1145/3719027.3765076.

[JOB+18] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, et al. “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning”. In: 2018 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2018, pp. 19–35. doi: 10.1109/SP.2018.00057.

[JSM+19] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. “PRADA: Protecting Against DNN Model Stealing Attacks”. In: 2019 IEEE European Symposium on Security and Privacy (EuroS&P). 2019, pp. 512–527. doi: 10.1109/EuroSP.2019.00044.

[KCM25] Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz. “Disparate Privacy Vulnerability: Targeted Attribute Inference Attacks and Defenses”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/kabir.

[KDD25] Torsten Krauß, Hamid Dashtbani, and Alexandra Dmitrienko. “TwinBreak: Jailbreaking LLM Security Alignments based on Twin Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/krauss.

[KKC25] Yujin Kang, Eunsun Kim, and Yoon-Sik Cho. “Can Personal Health Information Be Secured in LLM? Privacy Attack and Defense in the Medical Domain”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4199–4213. doi: 10.1145/3719027.3765105.

[KNT+25] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. “Invisible but Detected: Physical Adversarial Shadow Attack and Defense on LiDAR Object Detection”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/kobayashi.

[KSM+25] Seongho Keum, Dongwon Shin, Leo Marchyok, Sanghyun Hong, and Sooel Son. “Private Investigator: Extracting Personally Identifiable Information from Large Language Models Using Optimized Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/keum.

[LBW+18] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, et al. Understanding Membership Inferences on Well-Generalized Learning Models. 2018. arXiv: 1802.04889 [cs.CR].

[LDL+24] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, et al. Prompt Injection attack against LLM-integrated Applications. 2024. arXiv: 2306.05499 [cs.CR].

References (Part 5 of 10)

[LHS+25] Yang Lou, Haibo Hu, Qun Song, Qian Xu, Yi Zhu, et al. “Asymmetry Vulnerability and Physical Attacks on Online Map Construction for Autonomous Driving”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3251–3265. doi: 10.1145/3719027.3765092.

[LLM+25] Taifeng Liu, Yang Liu, Zhuo Ma, Tong Yang, Xinjing Liu, et al. “L-HAWK: A Controllable Physical Adversarial Patch Against a Long-Distance Target”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/l-hawk-a-controllable-physical-adversarial-patchagainst-a-long-distance-target/.

[LLW+25a] Hao Li, Zheng Li, Siyuan Wu, Yutong Ye, Min Zhang, et al. “Enhanced Label-Only Membership Inference Attacks with Fewer Queries”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/li-hao.

[LLW+25b] Xiao Li, Yue Li, Hao Wu, Yue Zhang, Kaidi Xu, et al. “Make a Feint to the East While Attacking in the West: Blinding LLM-Based Code Auditors with Flashboom Attacks”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 576–594. doi: 10.1109/SP61157.2025.00125.

[LMX+25] Shuofeng Liu, Mengyao Ma, Minhui Xue, and Guangdong Bai. “Modifier Unlocked: Jailbreaking Text-to-Image Models Through Prompts”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 355–372. doi: 10.1109/SP61157.2025.00242.

[LPC+25] Changjiang Li, Ren Pang, Bochuan Cao, Jinghui Chen, Fenglong Ma, et al. “Watch the Watchers! On the Security Risks of Robustness-Enhancing Diffusion Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/li-changjiang.

[LPH+25] Andrey Labunets, Nishit V. Pandya, Ashish Hooda, Xiaohan Fu, and Earlence Fernandes. “Fun-tuning: Characterizing the Vulnerability of Proprietary LLMs to Optimization-Based Prompt Injection Attacks via the Fine-Tuning Interface”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 411–429. doi: 10.1109/SP61157.2025.00121.

[LQS25] Chris S. Lin, Joyce Qu, and Gururaj Saileshwar. “GPUHammer: Rowhammer Attacks on GPU Memories are Practical”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/lin-shaopeng.

[LSS+23] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, et al. “Analyzing Leakage of Personally Identifiable Information in Language Models”. In: 2023 IEEE Symposium on Security and Privacy (SP). 2023, pp. 346–363. doi: 10.1109/SP46215.2023.10179300.

[LSZ+25] Harry Langford, Ilia Shumailov, Yiren Zhao, Robert Mullins, and Nicolas Papernot. “Architectural Neural Backdoors from First Principles”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 1657–1675. doi: 10.1109/SP61157.2025.00060.

[LWW+25] Zongjie Li, Daoyuan Wu, Shuai Wang, and Zhendong Su. “Differentiation-Based Extraction of Proprietary Data from Fine-Tuned LLMs”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3071–3085. doi: 10.1145/3719027.3744856.

[LWX+24] Shaofeng Li, Xinyu Wang, Minhui Xue, Haojin Zhu, Zhi Zhang, et al. “Yes, One-Bit-Flip Matters! Universal DNN Model Inference Depletion with Runtime Code Fault Injection”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1315–1330. url: https://www.usenix.org/conference/usenixsecurity24/presentation/li-shaofeng.

References (Part 6 of 10)

AISI Japan AI Safety Institute

[MGC22] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. “Property Inference from Poisoning”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1120—1137. doi: 10.1109/SP46214.2022.9833623.

[MMR+25] Kaleel Mahmood, Caleb Manicke, Ethan Rathbun, Aayushi Verma, Sohaib Ahmad, et al. “Busting the Paper Ballot: Voting Meets Adversarial Machine Learning”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3132—3146. doi: 10.1145/3719027.3744882.

[MZK+24] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, et al. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. 2024. arXiv: 2312.02119 [cs.LG]

[NFI+25] Milad Nasr, Yanick Fratantonio, Luca Invernizzi, Ange Albertini, Loua Farah, et al. “Evaluating the Robustness of a Production Malware Detection System to Transferable Adversarial Attacks”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4394—4408. doi: 10.1145/3719027.3765118.

[NMF+24] Najmeh Nazari, Hosein Mohammadi Makrani, Chongzhou Fang, Hossein Sayadi, Setareh Rafatirad, et al. “Forget and Rewire: Enhancing the Resilience of Transformer-based Models against Bit-Flip Attacks”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1349—1366. url: https://www.usenix.org/conference/usenixsecurity24/presentation/nazari.

[NPS+25] Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, et al. “Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 1245—1259. doi: 10.1145/3719027.3744840.

[NRB+25] Ali Naseh, Jaechul Roh, Eugene Bagdasarian, and Amir Houmansadr. “Backdooring Bias (B2) into Stable Diffusion Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/naseh.

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning”. In: 2019 IEEE Symposium on Security and Privacy (SP). 2019, pp. 739—753. doi: 10.1109/SP.2019.00065.

[NYW+25] Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, et al. “Rectifying Privacy and Efficacy Measurements in Machine Unlearning: A New Inference Attack Perspective”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/naderloui.

[OSF19a] Seong Joon Oh, Bernt Schiele, and Mario Fritz. “Towards Reverse-Engineering Black-Box Neural Networks”. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Cham: Springer International Publishing, 2019, pp. 121—144. doi: 10.1007/978-3-030-28954-6_7.

[OSF19b] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Knockoff Nets: Stealing Functionality of BlackBox Models”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, pp. 4949—4958. doi: 10.1109/CVPR.2019.00509.

[OVA+25] David Oygenblik, Abhinav Vemulapalli, Animesh Agrawal, Debopam Sanyal, Alexey Tumanov, et al. “VillainNet: Targeted Poisoning Attacks Against SuperNets Along the Accuracy-Latency Pareto Frontier”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2189—2203. doi: 10.1145/3719027.3765185.

[PEC+25] Rodrigo Pedro, Miguel E. Coimbra, Daniel Castro, Paulo Carreira, and Nuno Santos. “Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses”. In: 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 76—88. doi: 10.1109/ICSE55347.2025.00007.

References (Part 7 of 10)

[PHS+22] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, et al. “Red Teaming Language Models with Language Models”. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, Dec. 2022, pp. 3419–3448. doi: 10.18653/v1/2022.emnlp-main.225.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, et al. “Practical Black-Box Attacks against Machine Learning”. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ASIA CCS ’17. Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2017, pp. 506–519. doi: 10.1145/3052973.3053009.

[PW25] Yan Pang and Tianhao Wang. “Black-box Membership Inference Attacks against Fine-tuned Diffusion Models”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/black-box-membership-inference-attacks-against-fine-tuned-diffusion-models/.

[RCY+22] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. “DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1157–1174. doi: 10.1109/SP46214.2022.9833743.

[RSE25] Mark Russinovich, Ahmed Salem, and Ronen Eldan. “Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich.

[Sam23] Roman Samoilenko. “New Prompt Injection Attack on ChatGPT Web Version. Markdown Images can Steal Your Chat Data.” In: System Weakness (Mar. 2023). url: https://systemweakness.com/newprompt-injection-attack-on-chatgpt-web-version-ef717492c5c2.

[SCB+24] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. “"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models”. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS ’24. Salt Lake City, UT, USA, 2024, pp. 1671–1685. doi: 10.1145/3658644.3670388.

[Sch19] Oscar Schwartz. “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation”. In: IEEE Spectrum (Nov. 2019). Updated: Jan. 2024. url: https://spectrum.ieee.org/in- 2016- microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[SHJ+20] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods”. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 180–186. doi: 10.1145/3375627.3375830.

[SKK+25] Minkyoo Song, Hanna Kim, Jaehan Kim, Seungwon Shin, and Sooel Son. “Refusal Is Not an Option: Unlearning Safety Alignment of Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/song-minkyoo.

[SQB+24] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. “Prompt Stealing Attacks Against Text-to-Image Generation Models”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 5823–5840. url: https://www.usenix.org/conference/ usenixsecurity24/presentation/shen-xinyue.

[SRS17] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Machine Learning Models that Remember Too Much”. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. CCS ’17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 587–601. doi: 10.1145/3133956.3134077.

[SS20] Congzheng Song and Vitaly Shmatikov. “Overlearning Reveals Sensitive Attributes”. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. url: https://openreview.net/forum?id=SJeNz04tDS.

References (Part 8 of 10)

[SSS+17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks Against Machine Learning Models”. In: 2017 IEEE Symposium on Security and Privacy (SP). 2017, pp. 3—18. doi: 10.1109/SP.2017.41.

[SSS25] Avital Shafran, Roei Schuster, and Vitaly Shmatikov. “Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/shafran.

[SZB+21] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, et al. “Sponge Examples: Energy-Latency Attacks on Neural Networks”. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). 2021, pp. 212—231. doi: 10.1109/EuroSP51992.2021.00024.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, et al. “Intriguing properties of neural networks”. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2014. arXiv: 1312.6199 [cs.CV].

[TSS+25] Yicong Tan, Xinyue Shen, Yun Shen, Michael Backes, and Yang Zhang. “On the Effectiveness of Prompt Stealing Attacks on In-the-Wild Prompts”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 392—410. doi: 10.1109/SP61157.2025.00120.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. “Stealing Machine Learning Models via Prediction APIs”. In: 25th USENIX Security Symposium (USENIX Security 16). Austin, TX: USENIX Association, Aug. 2016, pp. 601—618. url: https://www.usenix.org/ conference/usenixsecurity16/technical-sessions/presentation/tramer.

[VMP25] Corban Villa, Shujaat Mirza, and Christina Popper. “Exposing the Guardrails: Reverse-Engineering and Jailbreaking Safety Filters in DALL·E Text-to-Image Pipelines”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/villa.

[WBH+25] Stanley Wu, Ronik Bhaskar, Anna Yoo Jeong Ha, Shawn Shan, Haitao Zheng, et al. “On the Feasibility of Poisoning Text-to-Image AI Models via Adversarial Mislabeling”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2848—2862. doi: 10.1145/3719027.3744845.

[WFK+19] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. “Universal Adversarial Triggers for Attacking and Analyzing NLP”. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China, Nov. 2019, pp. 2153—2162. doi: 10.18653/v1/D19-1221.

[WG18] Binghui Wang and Neil Zhenqiang Gong. “Stealing Hyperparameters in Machine Learning”. In: 2018 IEEE Symposium on Security and Privacy (SP). 2018, pp. 36—52. doi: 10.1109/SP.2018.00038.

[WHS23] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. “Jailbroken: How Does LLM Safety Training Fail?” In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 80079—80110. url: https : //proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790- Abstract-Conference.html.

[WLC+25] Haolin Wu, Chang Liu, Jing Chen, Ruiying Du, Kun He, et al. “When Translators Refuse to Translate: A Novel Attack to Speech Translation Systems”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wu-haolin.

References (Part 9 of 10)

AISI Japan AI Safety Institute

[WWC+25] Lijin Wang, Jingjing Wang, Tianshuo Cong, Xinlei He, Zhan Qin, et al. “From Purity to Peril: Backdooring Merged Models From “Harmless” Benign Components”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wang-lijin.

[WYB+25] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. “On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wu-yixin-generation.

[WZS+25] Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, et al. “Mirage in the Eyes: Hallucination Attack on Multi-modal Large Language Models with Only Attention Sink”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wang-yining.

[WZZ+25] Zihao Wang, Rui Zhu, Zhikun Zhang, Haixu Tang, and XiaoFeng Wang. “Rigging the Foundation: Manipulating Pre-training for Advanced Membership Inference Attacks”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 2509—2526. doi: 10.1109/SP61157.2025.00177.

[XC25] Qi Xia and Qian Chen. “AlphaDog: No-Box Camouflage Attacks via Alpha Channel Oversight”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025 February 24-28, 2025. San Diego, CA USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/alphadog-no-box-camouflage-attacks-via-alpha-channel-oversight/.

[XDT+25] Binyan Xu, Xilin Dai, Di Tang, and Kehuan Zhang. “One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3087—3101. doi: 10.1145/3719027.3744859.

[YGF+18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting”. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 2018, pp. 268—282. doi: 10.1109/CSF.2018.00027.

[YHG+25] Yuqin Yan, Wei Huang, Ilya Grishchenko, Gururaj Saileshwar, Aastha Mehta, et al. “Relocate-Vote: Using Sparsity Information to Exploit Ciphertext Side-Channels”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/yan-yuqin.

[YLD+25] Yuanyuan Yuan, Zhibo Liu, Sen Deng, Yanzuo Chen, Shuai Wang, et al. “CipherSteal: Stealing Input Data from TEE-Shielded Neural Networks with Ciphertext Side Channels”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 4136—4154. doi: 10.1109/SP61157.2025.00079.

[YLL+25] Yong Yang, Changjiang Li, Qingming Li, Oubo Ma, Haoyu Wang, et al. “PRSA: Prompt Stealing Attacks against Real-World Prompt Services”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/yang-yong.

[YSQ25] Kai Ye, Liangcai Su, and Chenxiong Qian. “ImportSnare: Directed ’Code Manual’ Hijacking in Retrieval-Augmented Code Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 335—349. doi: 10.1145/3719027.3765161.

[YYC+24] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, et al. “On the Vulnerability of Safety Alignment in Open-Access LLMs”. In: Findings of the Association for Computational Linguistics: ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9236—9260. doi: 10.18653/v1/2024.findings-acl.549.

References (Part 10 of 10)

[YZG+25] Xuejing Yuan, Jiangshan Zhang, Feng Guo, Kai Chen, XiaoFeng Wang, et al. “EvilHarmony: Stealthy Adversarial Attacks Against Black-Box Speech Recognition Systems”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 4569—4587. doi: 10.1109/SP61157.2025.00259.

[YZH+25] Dayong Ye, Tianqing Zhu, Feng He, Bo Liu, Minhui Xue, et al. “Cross-Modal Prompt Inversion: Unifying Threats to Text and Image Generative AI Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-inversion.

[YZL+25] Dayong Ye, Tianqing Zhu, Jiayang Li, Kun Gao, Bo Liu, et al. “Data Duplication: A Novel Multi-Purpose Attack Paradigm in Machine Unlearning”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-duplication.

[YZW+25] Dayong Ye, Tianqing Zhu, Shang Wang, Bo Liu, Leo Yu Zhang, et al. “Data-Free Model-Related Attacks: Unleashing the Potential of Generative AI”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-attacks.

[ZCI24] Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. “Effective Prompt Extraction from Language Models”. In: First Conference on Language Modeling. 2024. url: https://openreview.net/forum?id=0o95CVdNuz.

[ZCS+25] Ruofan Zhu, Ganhao Chen, Wenbo Shen, Xiaofei Xie, and Rui Chang. “My Model is Malware to You: Transforming AI Models into Malware by Abusing TensorFlow APIs”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 486—503. doi: 10.1109/SP61157.2025.00012.

[ZGW+25] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

[ZGY+25] Lan Zhang, Xinben Gao, Liuyi Yao, Jinke Song, and Yaliang Li. “Exploiting Task-Level Vulnerabilities: An Automatic Jailbreak Attack and Defense Benchmarking for LLMs”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zhang-lan.

[ZHK22] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. 2022. arXiv: 2209.10505 [cs.CL].

[ZWC+23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. 2023. arXiv: 2307.15043 [cs.CL].

[ZWZ+24] Jian Zhao, Shenao Wang, Yanjie Zhao, Xinyi Hou, Kailong Wang, et al. “Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs”. In: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. ASE ’24. Sacramento, CA, USA: Association for Computing Machinery, 2024, pp. 2087—2098. doi: 10.1145/3691620.3695271

[ZZM+25] Tingwei Zhang, Collin Zhang, John X. Morris, Eugene Bagdasarian, and Vitaly Shmatikov. “Self-interpreting Adversarial Images”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zhang-tingwei.

AISI

Japan AI Safety Institute

Japan AI Safety Institute
"Known Attacks and Their Impacts on AI Systems (2nd Edition)" (Apr. 2026)
https://aisi.go.jp/output/output_security/known_attacks_and_impacts/